

Copyright
by
Hye Sun You
2016

**The Dissertation Committee for Hye Sun You Certifies that this is the approved
version of the following dissertation:**

**Toward Interdisciplinary Science Learning: Development of an
Assessment for Interdisciplinary Understanding of ‘Carbon Cycling’**

Committee:

Jill A. Marshall, Supervisor

Mona Mehdy

Victor Sampson

Cesar Delgado

Diane L. Schallert

**Toward Interdisciplinary Science Learning: Development of an
Assessment for Interdisciplinary Understanding of ‘Carbon Cycling’**

by

Hye Sun You, B.S.; M.Ed.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2016

For Kyungun, my lovely husband:
Thank you for your never-ending love and support over these six long years while I
pursued my dreams.

Acknowledgments

There are many individuals who deserve recognition for helping to support me through the last six years in which I have been able to follow my passion for science education. Foremost, my husband, Kyungun, deserves a wealth of recognition for all of the support, encouragement, and love he has provided me with throughout graduate school. He has stood by me and sacrificed to help me accomplish my goals even when I know I was a challenge to live with. I could not ask for a better companion. I am indebted to my parents, Hyunwoo Yang and Byungil You for a lifetime of support of my ambitions, for fostering my interest in science as I was growing up. In addition to my parents, I am greatly appreciative and thankful for the large support and encouragement throughout the most difficult parts of graduate school of my mother- and father-in-law (Hyunkyung Park and Ingeol Kim). Your love and hope never allowed me to give up.

There are many individuals who have made this dissertation completed. First, I am extremely grateful to my academic advisor, Dr. Jill Marshall, for offering encouragement and guidance when needed. I could not have done this without you. Additionally, I would like to thank my previous academic advisor, Dr. Cesar Delgado, for his endless support when I struggled through my dissertation and other research. I have greatly appreciated him giving me the independence to expand upon and pursue my research interests throughout graduate school. My committee members have been supportive during my dissertation stage through their great comments. I would like to express my sincere appreciation to Dr. Mona Mehdy, Dr. Diane Schallert, and Dr. Victor Sampson.

Also, many friends from all aspects of my life have contributed to my success. Of those, Seoungyeon and Gina! Thank you for the friendship and your constant words of encouragement which helped keep me on track. Thank Moses and Jungsoo, for editing and formatting my dissertation and, for your help and support.

Lastly, even though my son, Gwidong (this is a fetal nickname) is not born yet, thank you so much for bearing up all difficulties during all dissertation processes with me. I love you so much!

Toward Interdisciplinary Science Learning: Development of an Assessment for Interdisciplinary Understanding of ‘Carbon Cycling’

Hye Sun You, Ph.D.

The University of Texas at Austin, 2016

Supervisor: Jill A. Marshall

This study aimed at developing and validating an assessment that measures interdisciplinary understanding for the topic carbon cycling. The impetus for this study is the recognition of assessment as “the ‘black hole’ of interdisciplinary education” in K-16 science education (Boix Mansilla, 2005, p. 18). There is no question that the complexity of natural systems and the corresponding scientific problems necessitates interdisciplinary understanding informed by multiple disciplinary backgrounds.

This study followed the construct-modeling framework for the interdisciplinary science assessment (ISA) design process (Wilson, 2005). A construct map for interdisciplinary understanding of carbon cycling was developed. Nine different subtopics within carbon cycling were determined based on content experts’ concept maps and analyses of the Next Generation Science Standards. Initial items were reviewed by content experts and piloted with students to establish content validity. Through the item revision process, a final version of the ISA was developed including 11 multiple-choice (MC) items and eight constructed response (CR) items. 454 students (9th grade to college seniors) were recruited and administered the ISA through the Qualtrics online environment. For the CR items scoring rubrics were developed and used to code student responses by a group of evaluators. Agreement between coders was greater than 90%, and analysis of scores indicated excellent inter-rater reliability. Item Response Theory (IRT) models, a two Parameter Logistic Model and a Generalized Partial Credit Model, provided evidence of the construct validity of the assessment items. All items reflected

unidimensional construct and local independency in the IRT analyses. All except one item were a good fit to the models. The misfit item was too easy for the range of student performance levels. Two items functioned differentially across gender, indicated a possible bias. The 19 items showed modest internal consistency (Cronbach's $\alpha = 0.782$). The findings suggest that the ISA is a promising and valid tool to assess interdisciplinary understanding in learning carbon cycling but the one misfit item and two DIF items merit further revision to strengthen the psychometric properties of the ISA. It is believed that the shift in the perspective of assessment towards interdisciplinary understanding enables science teachers to design their curriculum and instructional practices in a way that their students can learn how to connect one concept to another across different science disciplines, improving their scientific literacy.

Table of Contents

| | |
|--|----|
| Chapter 1: Introduction..... | 1 |
| 1.1. Background..... | 1 |
| 1.2. Rationale for the study..... | 3 |
| 1.3. Research questions | 5 |
| Chapter 2: Literature Review | 7 |
| 2.1. History of interdisciplinary curriculum | 8 |
| 2.2. Interdisciplinary learning and teaching in national standards for science education..... | 12 |
| 2.3. Different typologies of interdisciplinarity | 16 |
| 2.4. Framework of interdisciplinary science learning | 20 |
| Piagetian constructivism..... | 21 |
| Expert-novice. | 22 |
| Knowledge integration. | 26 |
| 2.5. Benefits of interdisciplinary learning | 28 |
| 2.6. Definition of interdisciplinary understanding | 30 |
| 2.7. Studies on interdisciplinary learning and teaching in science education | 31 |
| 2.8. Literature on assessments associated with interdisciplinary understanding in science education | 38 |
| 2.9. National and international tests and project-based assessments..... | 41 |
| 2.10. Assessment triangle | 46 |

| | |
|---|-----|
| 2.11. Construct-modeling | 48 |
| 2.12. Validity | 50 |
| Chapter 3: Methods | 56 |
| 3.1. Overview of the research design | 56 |
| 3.2. Participants and data collection | 58 |
| 3.3. Interdisciplinary understanding construct map | 59 |
| 3.4. Item design | 60 |
| 3.5. Outcome space (rubric development and validation process) | 65 |
| 3.6. Measurement model: two-parameter logistic model (2PLM) and generalized partial credit model (GPLM) | 70 |
| 3.7. Data analyses | 73 |
| Chapter 4: Results..... | 80 |
| 4.1. Development of the construct map | 80 |
| 4.2. Item design | 81 |
| 4.3. Inter-rater reliability | 92 |
| 4.4. Results of data analysis | 93 |
| Chapter 5: Discussion and Conclusions | 112 |
| 5.1. Implications of the interdisciplinary science assessment through a comparison with other literature | 112 |
| 5.2. Addressing construct validity | 114 |
| 5.3. Making inferences about interdisciplinary understanding scores..... | 116 |

| | | |
|---|---|-----|
| 5.4. | Cognitive process between disciplinary and interdisciplinary learning | 116 |
| 5.5. | Linking interdisciplinary science assessment to instruction | 118 |
| 5.6. | Directions for future research | 120 |
| 5.7. | Limitations | 121 |
| 5.8. | Conclusions | 122 |
| Appendixes | | 124 |
| Appendix A: Interdisciplinary science assessment (ISA) | | 124 |
| Appendix B: Scoring rubrics | | 136 |
| Appendix C: Item characteristic and item information curves | | 154 |
| References | | 159 |

List of Tables

| | |
|--|-----|
| Table 1. Percentage of TIMSS science assessment score points at grade 4 and 8 devoted to content and cognitive domains in 2011..... | 43 |
| Table 2. Demographic information. | 59 |
| Table 3. Proportion of experts whose endorsement is required to establish content validity beyond the .05 level of significance (Lynn, 1986). | 63 |
| Table 4. Holistic rubric on interdisciplinary understanding. | 67 |
| Table 5. Construct map of the interdisciplinary understanding of carbon cycling..... | 81 |
| Table 6. CVI ratings by experts on 20 items. | 91 |
| Table 7. Percentage agreement and intraclass correlation coefficients assessing inter-reliability..... | 93 |
| Table 8. Demographic information and descriptive statistics (N=454). | 95 |
| Table 9. CFA model fit indices of ISA-one factor model. | 96 |
| Table 10. CFA model fit indices of ISA-two factor model. | 96 |
| Table 11. S-X ² Item Level Diagnostic Statistics. | 99 |
| Table 12. 2PLM item parameters estimates, logit: $a\theta + c$ or $a(\theta - b)$ | 100 |
| Table 13. GPC model item parameter estimates, logit: $a[k(\theta - b) + \sum dk]$ | 101 |
| Table 14. DIF statistics for the ISA items. | 105 |
| Table 15. Point-Biserial correlations between individual items subscales..... | 107 |
| Table 16. Summary of Fit Statistics for the Models Tested. | 108 |
| Table 17. Descriptive statistics of four groups categorized by the number of course taken. | 109 |

| | |
|---|-----|
| Table 18. Descriptive statistics for grades..... | 110 |
| Table 19. Descriptive statistics of ISA theta scores for race. | 111 |

List of Figures

| | |
|--|-----|
| <i>Figure 1.</i> Characteristics of multidisciplinary, interdisciplinary and transdisciplinary. | 20 |
| <i>Figure 2.</i> Developmental sequences representing structural changes..... | 25 |
| <i>Figure 3.</i> Assessment triangle. | 48 |
| <i>Figure 4.</i> Four building blocks of construct modeling (Wilson, 2005)..... | 50 |
| <i>Figure 5.</i> Graphic display of the instrument development and validation process. | 57 |
| <i>Figure 6.</i> Three hypothesized structural models of the ISA. | 78 |
| <i>Figure 7.</i> Interdisciplinary connections in the NGSS | 82 |
| <i>Figure 8.</i> Concept map examples of the content experts. | 88 |
| <i>Figure 9.</i> One factor CFA model of the ISA..... | 97 |
| <i>Figure 10.</i> Scree plot of the ISA. | 98 |
| <i>Figure 11.</i> Item characteristic curve and item information curve for Item 5. | 102 |
| <i>Figure 12.</i> Item characteristic curve and item information curve for Item 2. | 102 |
| <i>Figure 13.</i> Facet map of the ISA. | 104 |
| <i>Figure 14.</i> Model of cognitive process of crosscutting concepts in NGSS..... | 118 |

Chapter 1: Introduction

1.1. Background

Some important themes pervade science, mathematics, and technology and appear over and over again, whether we are looking at an ancient civilization, the human body, or a comet. They are ideas that transcend disciplinary boundaries and prove fruitful in explanation, in theory, in observation, and in design.
(American Association for the Advancement of Science [AAAS], 1989, p.123)

Differentiation of natural science disciplines into the present disciplines such as physics, chemistry, and biology has a relatively short history of 200 years (Stichweh, 2003; Weingart, 2010). The specialization in science was due to the dramatic growth of the amount of scientific data and methods and the wish to handle the knowledge selectively in a certain specific discipline classification (Stichweh, 2003). Beyond the internal reasons, external social changes such as the invention of printing, population growth, and competition between disciplines played a crucial role in differentiating disciplines. The evolution of specialized science disciplines has exerted strong epistemological influences in shaping more differentiated science curriculum. For example, the differentiation of science curriculum affects selection and organization of learning objectives for students, as well as methods of teaching and assessment. However, starting in the early 1930s, the “unity of science movement” was initiated by natural scientists and philosophers of science, who argued that the closed boundaries of science disciplines are no longer the crucial frames for addressing real world problems (Neurath, 1996).

The educational paradigm shifts from purely disciplinary to interdisciplinary have been emerging since the mid-20th century. The need for an interdisciplinary perspective

is especially pertinent to science education because the natural phenomena studied are intrinsically interdisciplinary and real scientific issues are rarely confined to the artificial boundaries of academic disciplines. Rury (1996) stated that although science is highly fragmented and compartmentalized, the history of science disciplines illustrates that interdisciplinary pedagogy is a spontaneous process that is intrinsic to learning. Many contemporary scholars have also regarded the interdisciplinary approach as an essential alternative way for learning due to diverse educational benefits (Boix Mansilla, 2006; Clarke & Agne, 1997; Davis, 1995; Golding, 2009; Jacobs, 1989; Klein, 2002; Lattuca, Voigt, & Fath, 2004; Newell & Green, 1982). Newell and Green (1982) described the key benefits of interdisciplinary education as the integration of knowledge (connecting new knowledge to existing knowledge), deductive reasoning, and critical thinking. Similarly, as to the effectiveness of interdisciplinary learning, Lattuca et al. (2004) referred to outcomes such as

assisting students in developing complex understandings in particular subject areas, promoting the development of sophisticated views of knowledge, building students' capacity to recognize, evaluate, and use differing (multiple) perspectives, engaging student interest and increasing motivation; and enacting constructivist and active learning strategies. (p. 44)

Davis (1995) claimed that students in a complex society need to develop the ability to cope with multiple perspectives on issues and problems, and thus, interdisciplinary learning is especially well suited for encouraging complex views of knowledge among students.

In addition to the number of educators expressing appreciation of the significance of and advocate for interdisciplinary understanding across multiple disciplines, a variety

of US standards documents at the national level indicate the considerable amount of interest in and need for an interdisciplinary approach within science fields. The *Benchmarks for Science Literacy* (2009) suggested that science must be taught in a way to make connections with other science areas as well as other disciplines such as engineering. The *Next Generation Science Standards* (NGSS) (NGSS Lead States, 2013) highlighted that students need not only to develop insights and modes of thinking that are informed by a variety of disciplines, but also to form connections between fields of knowledge for desirable scientific literacy. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (“the framework”) (Council, 2012; National Research Council, 2012) emphasized disciplinary core ideas and meaningful connections of the disciplinary ideas across multiple scientific contexts through crosscutting concepts (CCCs) (e.g., energy and matter). *Developing Assessments for the Next Generation Science Standards* (National Research Council, 2014) asserted the need for new modes of assessment that reflect scientific ideas from multiple disciplinary perspectives. This trend towards interdisciplinary education imposes high demands on the development of a new system of assessment to measure interdisciplinary understanding in science.

In this regard, this study attempts to develop the interdisciplinary assessment with connections to chemistry, physics, geoscience and biology, providing insight for the process of developing reliable and validated items.

1.2. Rationale for the study

Disciplinary-based learning is still the norm in today's high school and college classrooms, even though many science topics in secondary and higher education are highly interdisciplinary across science disciplines. Traditional science assessments in national and international tests rely heavily on disciplinary-based items that require separate elements of scientific information. In this situation, students are prevented from discovering and creating links between any relevant science subjects, which eventually leads to poor interdisciplinary understanding of issues in science. The philosophy of the study is not saying that the disciplinary learning and interdisciplinary learning are mutually exclusive. Rather, it recognizes that each discipline is also important for interdisciplinary understanding. Approaching the boundary of a discipline can be a precursor to an interdisciplinary approach and students need a base of domain knowledge before integration of the knowledge. As a result, the interdisciplinary learning focuses neither blurs nor erases the disciplines; rather, it elucidates the usefulness of their distinctions by clarifying their universal connections (Metz, 1995).

Interdisciplinary learning has been promoted at all levels of education but there has been insufficient empirical research in the area of assessing students' interdisciplinary understanding (Shen, Liu, & Sung, 2014). Particularly, very little research has focused on interdisciplinary assessment in secondary schools. This situation calls for more authentic assessments that emphasize interdisciplinary understanding. This study suggests a new starting point for developing and validating an interdisciplinary assessment in science. The interdisciplinary assessment could provide a powerful impetus to curriculum and instruction, stimulating changes in curriculum policy and guiding the

professional development of teachers. Furthermore, the interdisciplinary assessment focusing on science and its conceptual framework can be modeled as a precursor to be applied and even expand the areas of technology, engineering, and mathematics.

1.3. Research questions

This study is motivated by the desire to assess students' interdisciplinary understanding on carbon cycling, along with the perspective of the NGSS and framework focusing on the ability to integrate content knowledge across science disciplines.

Assessment on interdisciplinary understanding is expected to shed some light on an important issue for implementing interdisciplinary learning and teaching. For purposes of this research, a validated and reliable assessment instrument will be developed using “construct modeling” (Shen et al., 2014; Wilson, 2005). This study aims to (a) develop a reliable and valid assessment instrument to measure the degree of high school and college students' interdisciplinary understanding across the disciplines of science, in the context of carbon cycling, (b) measure their interdisciplinary understanding, using the assessment tool developed, and (c) evaluate psychometric properties of student responses obtained from the assessment to establish the construct validity using Item Response Theory.

The guiding research questions for this study are as follows:

- 1) How valid and reliable is the developed assessment for measuring interdisciplinary understanding of carbon cycling? In order to answer this question, several sub-questions need answering to provide a variety of supporting evidence of construct validity and they are listed below.
 - a. To what extent do the sources of content validity index (CVI) evidence

support inferences about the items?

- b. What are the characteristics of the interdisciplinary science assessment items (unidimensionality, local independence, item fit, internal consistency etc.)?
 - c. How are the items on the instrument separated by difficulty and discrimination?
 - d. Do items on the instrument function differently across gender? (Does differential item functioning (DIF) occur?)
 - e. What proposed theoretical model best represents the internal structure of the instrument?
- 2) To what extent does the development process of scoring rubrics support inferences about the validity and inter-rater reliability of the scoring rubric developed for constructed response items?
- 3) How are the various characteristics of students (e.g., grade, gender, race, and the number of science courses taken) associated with their assessment performance?

Chapter 2: Literature Review

The research questions proposed in the first chapter are centered upon the development of assessment items for interdisciplinary understanding. Before investigating this, in this chapter, I delved into some fundamental understandings on the history of interdisciplinary learning integrated with science curriculum and the current trends in science education. Sections 2.1 and 2.2 illustrate a brief history of curriculum integration including the current paradigm change toward an integrated curriculum in U.S. education. Section 2.3 summarizes the definitions of ‘interdisciplinary’ and analogous terms to gain a consistent understanding of what ‘interdisciplinary’ is and how ‘interdisciplinary’ differs from other similar terms. Section 2.4 discusses learning theories and their related research studies that can support the rationale and justifications for interdisciplinary learning and section 2.5 emphasizes the importance of the interdisciplinary assessment through the process of learning and teaching. Section 2.6 shows the theoretical definition of students’ interdisciplinary understanding and thinking, which provides a framework for constructing an operationalized definition of ‘interdisciplinary understanding’. Section 2.7 summarizes the previous literature on interdisciplinary learning and teaching while section 2.8 focuses on reviewing interdisciplinary assessment studies in science education areas at both secondary and college levels. Section 2.9 reviews the possible national and international tests to see if they could show students’ interdisciplinary understanding across various disciplines as well as examples of existing assessments developed for curricular projects or research studies. Section 2.10 shows the ‘assessment triangle’ to describe assessment as a process

of reasoning from evidence and section 2.11 reviews a construct-centered assessment design, ‘construct-modeling’, which will be used in this study. Lastly, section 2.12 describes the types of validity and their definitions based on Messick’s framework (1995).

2.1. History of interdisciplinary curriculum

The historical perspective of interdisciplinary curriculum provides evidence that educational reformers long before our time have advocated integrated learning in educational systems within the flow of American education history (Beane, 1995; Chandramohan & Fallows, 2009; Kliebard, 2004). The “interdisciplinary” term was first introduced in curricular contexts in the 1920s and the term has been used for around 100 years (Klein, 1990); however, the concept of ‘interdisciplinary’ had existed even before the emergence of the term. Confucius and Socrates were the first scholars who presented the root of an interdisciplinary approach to learning (Henson, 2003). Plato advocated a synthesis between knowledge and unified science (Klein, 1990). Aristotle also was a philosopher who emphasized the innate ability to gather all kinds of knowledge and organize them to form broader or innovative concepts. The Aristotelian belief in the unity of knowledge is frequently cited as a rationale for interdisciplinary education. During the 18th century, Jean Rousseau expanded the interdisciplinary concept by applying it to a branch of child-centered education (Henson, 2003). Actual curriculum integration for interdisciplinary learning began in the late 1800s with the Herbartian movement (Drake & Burns, 2004). The curriculum, before the Herbartian movement, was segmented and isolated by subjects, discouraging any connections or relationships between them. To

rectify the system, Tuiskon Ziller, a disciple of Herbart proposed the idea of correlating disconnected subject areas around specific themes, sometimes referred to as “integration of studies” (Klein, 2002). The basic idea in “correlation” was the arrangement of subjects in a curriculum in such a way to organize the course of studies so that the matter of the different branches was simultaneously treated. Along with the development of the concept pertaining to the correlation of subjects and unity of learning, Herbartian scholars developed five steps (i.e., 1. Preparation, 2. Clear presentation of ideas, 3. Association of ideas, 4. Classification of ideas, 5. Application of ideas) in the construction of knowledge. These five steps, later combined with Dewey’s problem method, became the basis for the concept of the integration curriculum. The Herbartians’ key idea, concerning the variety of branches of school disciplines, could be correlated with the development of the interdisciplinary concept. This idea reached its peak around 1980s and has become the basis for the concept of interdisciplinary curriculum within modern day American education (Wraga, 1996).

In addition, the underlying concept of interdisciplinary learning can be traced in the history of the progressive education movement in the U.S. during the first half of the twentieth century. This movement has been divided into two competing groups: administrative progressivism versus pedagogical progressivism. Administrative progressives mainly focused on the scientific and differentiated curriculum, and acknowledged the existence of developmental differences in children of the same age groups (Labaree, 2005). They also emphasized the outcomes of a curriculum in that children’s roles were only to meet the needs of the society (Labaree, 2005). However,

today's interdisciplinary learning is much closer to pedagogical progressivism. The basic philosophy of pedagogical progressive highlights the Herbartians' idea that curriculum and instruction have to be "child-centered", which can be achieved by integrating disciplines that correlate with socially relevant themes (Labaree, 2005). Two important components in pedagogical progressivism are developmentalism and holistic learning (Hirsch, 1996). If learning is natural, then teaching needs to acclimate to the learner, which means that a careful selection of subject topics and skill levels has to be coordinated in order to steadily follow a student's pace of development.

"Developmentally appropriate" practices and curricula are fundamental in pedagogical philosophy. The holistic learning of pedagogical progressivism states that authentic natural learning only occurs in a holistic manner, where several realms of skill and knowledge are integrated as units, topics, and projects rather than being taught as separate subjects. There were a number of prominent figures spearheading and representing pedagogical progressivism, including G. Stanley Hall, William Kilpatrick, and Harold Rugg. Out of them all, Dewey is a pioneer who led the pedagogical progressive education movement and provided insights into major implications for the current interdisciplinary learning. Dewey (1938) advocated a child-centered learning environment where the educational experiences of children involved the principles of "continuity" and "interaction." He believed that a curriculum based on personal experiences leads to natural connections between prior knowledge and learning of new material. In contrast, intentionally separated subjects may prevent children from finding and establishing the relationships among the relevant subjects. Gehrke (1998) identified

two popular periods of interdisciplinary curriculum history during the 20th century: the progressive era of the 1920s and the 1930s, and the open education movement period of the 1960s and the early 1970s. However, the back-to-basics movement, which began in the 1970s, made the integrated science curriculum movement slow dramatically. One of the main reasons that the movement of curriculum integration faded away was the launching of Sputnik (1957). The post-Sputnik reforms of the federal government called for a more academic system with a utilitarian approach to increase the efficiency of learning (Schramm, 2001). A change in the recognition of schools' roles and responsibilities for society led to substantial support for a discipline-based curriculum. The proponents wanted a steady control of practices related to teaching and learning in schools in order to obtain better performance from the students. However, during the 1980s, educators tried to find the balance between specialization and integration. Since the 1990s, scholars have paid close attention in designing and managing interdisciplinary curricular and associated research projects, practical and philosophical consequences of relationships between particular disciplines, and the nature of interdisciplinary theories and methods (Klein, 1990). By doing so, the number of published journals and books greatly increased, showing the importance of interdisciplinary learning and teaching and suggesting better ways to go at it during the 1990s (Gehrke, 1998).

Beyond the perspectives of the progressive movement, postmodernism has significantly influenced the interdisciplinary curriculum development (Villaverde, 2003). Postmodernism solidified the paradigm shift in theoretical frameworks and provided a new philosophy in reconceptualizing curriculum. Postmodern curriculum uses techniques

such as the utilization of varying disciplines in understanding and learning a new phenomenon in order for us to expand and interpret knowledge (Villaverde, 2003). Doll (1993) stated that the disciplinary model just leads to a linear, sequential, easily quantifiable system, which is not satisfactory to postmodernist educators. By contrast, a postmodernist curriculum seeks comprehensive and integrated knowledge to maximize students' learning, moving further away from the modernist viewpoint.

In conclusion, the pedagogical progressive movement and postmodernism were influential forces that shaped interdisciplinary curriculum in the modern American schooling system. With the growing recognition of the importance of interdisciplinary learning and teaching, many reformers and researchers today are attempting to imbue ideas of interdisciplinary learning and teaching into the current education system (Boix Mansilla & Duraisingh, 2007; Boix Mansilla, Miller, & Gardner, 2000; Clarke & Agne, 1997; Golding, 2009; Jacobs, 1989; Klein, 2002).

2.2. Interdisciplinary learning and teaching in national standards for science education

A variety of U.S. standards have already recognized the need and importance of interdisciplinary approaches to science learning a half-century ago. *Theory Into Action* (National Science Teachers Association, 1964) by the NSTA Curriculum Committee showed the importance of common themes in science subjects. After the back-to-basics movement in education, in the late 1980s, ideas of integrating the science subjects began reappearing in the California Science Framework, which stated, “in order for science to

be a philosophical discipline and not merely a collection of facts, there must be thematic connection and integration” (California Department of Education, 1990, p. 2).

In 1989 a strategy to reunite the sciences, Scope, Sequence, and Coordination (SS&C), was initiated by the National Science Teacher Association (McComas & Wang, 1998). The SS&C led interdisciplinary science instruction and introduced the ideas of the interdisciplinary orientation. The teaching standards for grades K-12 published by the National Science Teachers Association (1998) revealed the influence of integrated curriculum instruction. The *Benchmarks for Science Literacy* (AAAS, 2009) highlighted an inquiry- or process-based science curriculum and suggested that science must be taught in a way that makes connections between the sciences and other areas because science research occurs at the interface of various disciplines. The premise of ‘Project 2061’ lies in the importance of the interdisciplinary perspective that students should be equipped with interconnected knowledge across disciplines. It asserted that students' actual learning experience could occur in totally integrated contexts. For example, understanding the scientific explanation for *the evolution of life* depends on precursor knowledge of the physical sciences, Earth science, and some common themes that cut across disciplines. The National Science Education Standards (NSES) (National Research Council, 1996) stated, “Curricula often will integrate topics from different subject-matter areas such as life and physical sciences-from different content standards-such as life sciences and science in personal and social perspectives” (p. 23). The NSES (NRC, 1996) provided a framework called "Unifying Concepts and Processes" where boundaries between traditional subject matters are collapsed. Students are encouraged to be

productive and have an insightful way of thinking about connections made through *unifying concepts* to explain the natural and designed world. Unifying concepts and processes include 1) Systems, order, and organization, 2) Evidence, models, and explanation, 3) Change, constancy, and measurement, 4) Evolution and equilibrium, and 5) Form and function. The *Science Framework for the 2009 National Assessment of Educational Progress* (National Assessment Governing Board (2008) argued that one of the major differences between the 1996–2005 and 2009 NAEP science frameworks is the employment of crosscutting content.

In a similar vein, the concepts of "crosscutting ideas" in the National Science Teachers Association's Science Anchors Project (NSTA, 2010), and "unifying concepts" in *Science: College Board Standards for College Success* (College Board, 2009) were proposed. Recently, the NGSS (Lead States, 2013), and the Framework (NRC, 2012) also emphasized a conceptual shift towards crosscutting concepts (CCCs). The CCCs imply that the nature of science is not separate elements but intertwined aspects of knowledge and understanding that could be defined as the themes that bridge physical, life and Earth/space sciences, and engineering. The NGSS and Framework identified the following seven overarching CCCs.

1. Patterns. Observed patterns of forms and events guide organization and classification, and they prompt questions about relationships and the factors that influence them.
2. Cause and effect: Mechanism and explanation. Events have causes, sometimes simple, sometimes multifaceted. A major activity of science is investigating and explaining causal relationships and the mechanisms by which they are mediated. Such mechanisms

can then be tested across given contexts and used to predict and explain events in newer contexts.

3. Scale, proportion, and quantity. In considering phenomena, it is critical to recognize what is relevant at different measures of size, time, and energy and to recognize how changes in scale, proportion, or quantity affect a system's structure or performance.

4. Systems and system models. Defining the system under study—specifying its boundaries and making explicit a model of that system—provides tools for understanding and testing ideas that are applicable throughout science and engineering.

5. Energy and matter: Flows, cycles, and conservation. Tracking fluxes of energy and matter into, out of, and within systems helps one understand the systems' possibilities and limitations.

6. Structure and function. The way in which an object or living thing is shaped and its substructure determines many of its properties and functions.

7. Stability and change. For natural and built systems alike, conditions of stability and determinants of rates of change or evolution of a system are critical elements of study.

(p.84)

The CCCs have value because they provide students with an organizational schema and a coherent scientific view, which gradually helps the application practices across all domains of science (NRC, 2012, p.233). For example, “energy and matter” as one of the CCCs has powerful applications in physics, chemistry, mathematics, biology, engineering, etc., If students draw on ideas of how bodies get energy and matter out of foods, the concept of conservation of matter and transformation of energy, based on

physics and chemical reactions in a biological context, provides many opportunities for students to achieve interdisciplinary understanding.

The report *Developing Assessments for the Next Generation Science Standards* (NRC, 2014, xi) asserted, “the new K-12 framework makes it clear that such tools, reflecting new modes of assessment designed to measure the integrated learning it envisions, will be essential.”

2.3. Different typologies of interdisciplinarity

Despite a substantial body of work on the concept of interdisciplinarity, from the 1960s, there is still no consensus on how to define the typologies representing interdisciplinarity. The original OECD definition of ‘interdisciplinarity’ was relatively broad, ranging from “ simple communication of ideas to the mutual integration of organizing concepts, methodology, procedures, epistemology, terminology, data” (OECD, 1972, p. 25). Klein and Newell (1997) defined interdisciplinarity in detail as “a process of answering a question, solving a problem, or addressing a topic that is too broad or complex to be dealt with adequately by a single discipline and drawing on disciplinary perspectives and integrating their insights by producing a more comprehensive understanding.” (p. 2)

The five most widely used terms in typologies of interdisciplinarity are: ‘multidisciplinary’, ‘cross-disciplinary’, ‘interdisciplinary’, ‘pluridisciplinary’, and ‘transdisciplinary’. Since ‘interdisciplinary’ in research or/and practical contexts is interchangeably used in conjunction with other typologies, in order to have a more complete understanding of the meaning for ‘interdisciplinary’ the distinction between

interdisciplinary and other terms is necessary. The classification among typologies ending with ‘disciplinary’ is characterized by a terminological hierarchy in terms of the degree of tightness or looseness of the integration between bodies of knowledge (Klein, 1990; Kockelmans, 1979). The simplest form of interdisciplinarity between disciplines is multidisciplinary. It requires two or more disciplines that do not necessitate the integration/coherence/synthesis of knowledge, but it tends to present disciplines as non-interactive parallels (Chynoweth, 2009; Klein & Newell, 1997). Kockelmans (1979) stated that ‘multidisciplinarity does not have connections at all between disciplines involved. In multidisciplinary, students may have learned more than one subject simultaneously or in sequence without cognitive interaction between the subjects. While multidisciplinary indicates juxtaposition of various disciplines with no apparent connection between them (e.g. music + science + history), “pluridisciplinary” implies some juxtaposition of various disciplines, but the disciplines are assumed to be more or less related (e.g., mathematics + physics) (OECD, 1972). Cross-disciplinary has been used in the past to define a particular type of multidisciplinary (Stark & Lattuca, 1997; Stock & Burton, 2011). In cross-disciplinary learning, one discipline is used in the assistance of another, only through boundaries rather than integration. It suggests that boundaries are simply crossed rather than integrated; thus, crossdisciplinarity is a weaker form of the interdisciplinarity (Stock & Burton, 2011). Meeth (1978) described cross-disciplinarity as the next level after positioning disciplines in various levels on an integration pyramid that views one discipline from the perspective of another.

The word “interdisciplinary” stems from the Latin preposition *inter*, meaning

“between, among, in the midst,” or “derived from two or more” disciplines, meaning “of or relating to a particular field of study” or specialization (Stember, 1991, p. 4). The definitions and characteristics of “interdisciplinary” shown in a large body of literature varied depending on the context. Klein (1990) showed that the term “interdisciplinary” is used to describe close relations among academic disciplines, such as biology and ecology. The definition of interdisciplinary proposed by the National Academy of Sciences is the “integrat[ion] [of] information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice” (National Academy of Sciences, 2004, p. 26). Lederman and Niess (1997, p. 57) argued that “interdisciplinary” takes the perspectives of various disciplines, but it does not deny the existence of disciplines which serve to compartmentalize knowledge into separate units. Thus, the term “interdisciplinary” emphasizes the connections or interaction between subject matter, but there remains perceived value in the unique characteristics and distinctions among the various disciplines. These authors have two basic metaphors using the concepts derived from chemistry: mixture versus compound. While the characteristic of interdisciplinary maintaining disciplines’ identity is more of a mixture, the perception of integration reflects a seamless boundary among disciplines, similar to a compound in the chemistry analogy (Lederman & Niess, 1997). There is another metaphor about interdisciplinary, which is “bridge building” and preserving firm disciplines.

“Interdisciplinary” represents a radical state of reconstruction among the disciplines that parallels with the concept of “ integration” (Klein, 1990).

“Transdisciplinarity” is the most recent one of the five main concepts, going back to the early 1970’s and evolving from them (Balsiger, 2004). Miller (1982) explained that the “transdisciplinary approach is a conceptual framework that transcends the narrow scope of disciplinary world views, encompassing several parts of material handled separately by specialized disciplines” (p. 21). Transdisciplinarity may be the most desirable and yet difficult form of interdisciplinarity to attain (Stock & Burton, 2011). This form of interdisciplinarity draws on expertise from a wide range of disciplines that eventually produces the unity of intellectual frameworks that surpass disciplinary perspectives (Choi & Pak, 2006). Transdisciplinarity is concerned with humanistic issues such as pollution that transcends the narrow scope of disciplinary worldviews as well as synthetic theories like structuralism, Marxism, political science, general system theory and feminism (Klein, 1990). In the view of Jantsch (1947), transdisciplinarity indicates the ultimate degree of integration, where integrated parts of existing disciplines are formed into a new discipline, which in turn facilitates the enhancement of epistemologies.

| | Synthesize new disciplines and theory | Problem solving focus | Iterative research process | Involve multiple disciplines | Involve stakeholders in research process | Knowledge sharing between disciplines | Thematically based |
|-------------------|---------------------------------------|-----------------------|----------------------------|------------------------------|--|---------------------------------------|--------------------|
| Multidisciplinary | | | | | | | |
| Interdisciplinary | | | | | | | |
| Transdisciplinary | | | | | | | |

Figure 1. Characteristics of multidisciplinary, interdisciplinary and transdisciplinary. Adapted from only a part of original content of “Defining Terms for Integrated (Multi-Inter-Trans-Disciplinary) Sustainability Research by P. Stock, and R. J. F. Burton, 2011, Sustainability, 3, p.1101.

Stock and Burton (2011) displayed major differences in the nature of the three main typologies (see Figure 1). A filled box means that the component can be regarded as a characteristic of the type whereas an empty box indicates a consensus that it does not include the component, and a half-full box shows that there is some degree of argument as to whether this component is contained or not. As shown in Figure 1, interdisciplinary is similar to transdisciplinary. In fact, the only key differences between the two are that the transdisciplinary approach aims to synthesize new disciplines and generate a new theory whereas this is not an objective for interdisciplinarity. The boundaries between interdisciplinary and transdisciplinary projects are thus diffuse and dependent more on a subjective judgment on the level of holism applied than on the presence of clear boundary markers.

2.4. Framework of interdisciplinary science learning

Details of the theoretical perspective of cognitive constructivism, knowledge (or conceptual) integration, and novice-expert theory provide a supportive argument for and theoretical foundation on how students develop interdisciplinary understanding (Foss & Pinchback, 1998).

Piagetian constructivism. Constructivism provides a perspective on how children learn. In particular, Piagetian constructivism focuses on structures and use of knowledge, which consists of interconnected concepts, and emphasizes that learning is not simply the accumulation of new knowledge but it involves inducing the restructuring of learners' knowledge structures (Marzano, 1991). Piaget (1978) used concepts of schemas (i.e., the organization of information), assimilation, and accommodation to explain cognitive development and learning processes. For Piaget, "assimilation is the integration of external elements into evolving or completed structures" (1970, p. 706). Accommodation is "any modification of an assimilatory schema or structure by the elements it assimilates" (1970, p. 708). Whereas "assimilation is necessary in that it assures the continuity of structures and the integration of new elements to these structures" (1970, p. 707), accommodation is necessary to permit transformation of the existing schemas or creating new ones to provide a better fit for new knowledge. "Although structures or schemata can be changed and adapted, prior formulations are never destroyed or eliminated, and what was previously known remains with some improvement on the quality of knowledge." (Liu, Lee, & Linn, 2010, p. 25). The motivation for learning or learners to enter "new territory" comes from contradictions to their understandings (disequilibrium in Piaget's terms), which causes dissonance that necessitates a looping

cycle of interactions between assimilation, accommodation, and equilibrium that accommodates new information. The disequilibrium then leads to a qualitative change to one's thought processes, and, ultimately, to a new way of learning.

The framework of Piagetian constructivism provides principles for the interdisciplinary learning. For integrating knowledge, students are encouraged to restructure their existing knowledge domains from disciplines. As interdisciplinary learning occurs, increasingly well-structured and qualitatively different organizations of knowledge develop, which enables students to see the “big picture”, not a single disconnected piece of the larger puzzle. In addition, interdisciplinary education makes students aware of conflicts and inconsistencies in their thinking, generating disequilibrium. Piaget (1964) argued that disequilibrium is a requisite for cognitive development. Interdisciplinary education could provide learners with an environment that creates certain disequilibria where they can feel free to explore other areas and accommodate prior disconnected knowledge and create meaningful connections.

For the development of students' internal cognitive structures perceiving the entire context of knowledge, rather than only requiring disconnected basic information and recall of knowledge as in the current education system, new curriculum, instruction, and assessment systems incorporating the interdisciplinary approach will be needed.

Expert-novice. Within the expert-novice paradigm, numerous previous studies have attempted to identify the experts' characteristics in terms of a specific domain and problem solving (Chi & Bassok, 1989; Chi & Ceci, 1987; Chi, E, & Robin, 1988; Chi, Glaser, & Farr, 1988; Collins & Evans, 2007; Ericsson, Charness, Feltovich, & Hoffman,

2006; Ericsson, Nandagopal, & Roring, 2009; Kuchinke, 1997). Those studies have shown that experts possess more extensive and organized knowledge, making them more efficient in perceiving meaningful patterns, manipulating relevant information, and enabling them to perform excellently in practices compared to novices. Thus, experts solve a problem faster and more accurately, using knowledge structures that are more organized and easily accessible to them, than novices do (Bransford, Brown, & Cocking, 2000; Lehrer & Schauble, 2006).

Understanding the differences in cognitive processes between experts and novices provides a basis of recognizing the nature of interdisciplinary learning. Experts tend to find core concepts and central theoretical constructs in the cohesive framework of related concepts, and then further transfer them from one domain to another one in order to solve problems that are related with the given concept. Novices, on the other hand, tend to possess shallow concepts and isolate them as separate factual knowledge, which prevents them from comprehending or solving complex problems with an interdisciplinary approach.

According to the schema theory suggested by Sweller, Van Merriënboer, and Paas (1998), a complex schema is constructed by incorporating a large number of the interacting elements into a single element in long-term memory. Schema construction is enabled through the merging of lower level schemas into one higher-level schema, which plays a critical role in reducing working memory load. However, all learners are not on the same level in the process of schema construction. A main difference between experts and novices is that the experts have a wider range of existing knowledge in their long-

term memory in comparison with novices, which causes differences in the cognitive construction. Experts are also superior to novices in terms of making inferences on best-fitting new knowledge into existing knowledge clusters (Chi & Ceci, 1987). The ability allows experts to better perceive a grouped, meaningful pattern of the information and acquire more thematic knowledge. For example, Simon and Chase's study (1973) showed that expert chess players could not only identify isolated patterns, but also perceive an integrated configuration of chess piece positions. In contrast, novice chess players did not construct interconnected links. In the learning process as well, the knowledge of novices is limited and insufficient in understanding given core concepts and novices tends to have a lower sensitivity in recognizing the relationships between patterns and hierarchical classification of knowledge (Bransford et al., 2000). Chi and Ceci (1987, adapted from Keil, 1981) represented the gradually changing structural levels that show how cognitive development is processed in long-term memory. Specifically, the concept of "super links," which captures the developmental differences in general learning, can be applied in order to explain a process of development on interdisciplinary learning. As shown on the left diagram in Figure 2, as learning processes proceed with time, cognitive structures are changed qualitatively and quantitatively in ways in which the disconnected knowledge components have a "local coherence", which in turn provides "super links" between each module of knowledge. These coherently developed structures of knowledge with super links allow learners to see or understand the entire knowledge structure.

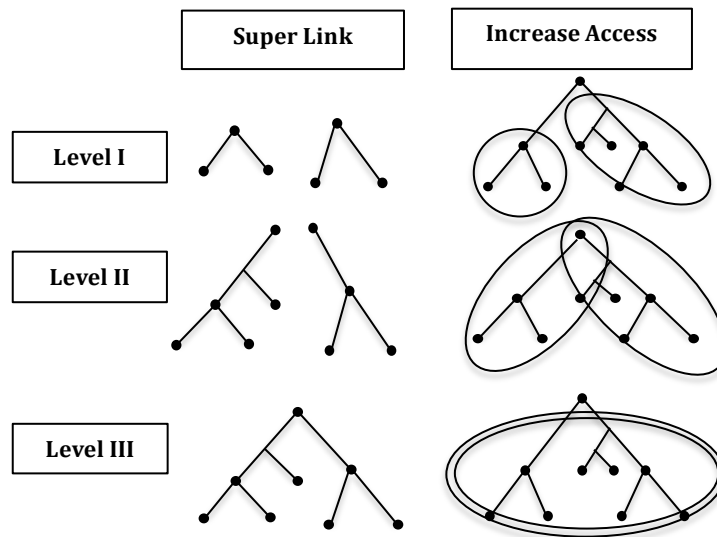


Figure 2. Developmental sequences representing structural changes. “Content knowledge: Its role, representation and restructuring in memory development,” by M. T. Chi, and S. J. Ceci, 1987, *Advances in Child Development and Behavior*, 20, p. 131. Copyright 1987 by the New York Academic Press. Reprinted with permission.

The second example of developmental structure changes illustrates Rozin’s (1976) theory of access. Even though learners have an entire and relevant amount of knowledge in long-term memory, there might be differences in the ability to access a wider range of the knowledge structure between novices and experts. The structure on the right in Figure 2 shows the obvious differences in the availability range of knowledge bases that are accessible. Some children can access some overlapping knowledge, while other children can access the entire knowledge base to solve a specific task. The above two examples of knowledge developmental structure provide key principles of experts’ knowledge and potential implications for both general and interdisciplinary learning. Particularly, these developmental structure diagrams offer an explanation for how

learners create strong relationships of a particular discipline to other disciplines in a more efficient manner (Ivanitskaya, Clark, Montgomery, & Primeau, 2002).

Knowledge integration. The knowledge integration (KI) theory provides a rationale in guiding interdisciplinary learning. The KI theory reflects a constructivist view of learning (Bransford et al., 2000; M. C. Linn, 2006). In the KI process, learning occurs when eliciting students' prior knowledge, adding new ideas that are potentially more powerful than existing ideas, comparing and contrasting between the new and old ideas, and developing criteria for appropriate applications of the new ideas, inevitably resulting in a more coherent understanding of science (M. C. Linn, 2006; Liu et al., 2010). M. C. Linn, Slotta, Terashima, Stone, and Madhok (2010) conceptualized the five dynamic processes of the KI as follows.

- Eliciting ideas. Learners elicit their own prior ideas, backgrounds and experiences, which enables them to create relevant connections to new ideas from already existing ideas in a learning context. For example, in a curriculum focused on the design of fuels, teaching can prompt students to elicit their existing observations and everyday ideas about energy and chemical reactions. Many studies have shown the benefits of eliciting ideas, in that students can develop a repertoire of ideas about scientific phenomena on the basis of their observations, experiences, and intellectual efforts.
- Adding new ideas. Learning environments traditionally aim to add ideas through some kind of learning activity, which allows learners to explore the relationships

among all their existing and new ideas and eventually form connections between those two.

- Distinguishing ideas. After adding ideas, students are required to carefully distinguish productive ideas from unproductive ones to connect scientifically relevant and normative ideas.
- Sorting out ideas. Students need opportunities to prioritize the numerous, often contradictory, existing ideas and sort out the various connections among the ideas to develop a coherent understanding of the subject.
- Developing criteria. Students need to develop criteria for the relationships between ideas. The criteria encourage students to coordinate productive ideas of target phenomena and demonstrate coherent and durable scientific understanding.

(p.5)

Shen, Liu, and Sung (2014) considered three special processes in interdisciplinary knowledge integration: translation, transfer, and transformation. The translation process involves specialized terminologies and jargon developed within each discipline that should be interpreted differently in other disciplines. Transfer refers to the process where students apply explanatory models and concepts learned from one disciplinary context to another. Transformation indicates the potential to apply explanatory models and concepts learned from one discipline to conceptually transform a system typically considered in a different discipline into a new system. The KI process implies that refining students' knowledge is a crucial step in drawing on their interdisciplinary understanding. Only

focusing on adding new ideas in classroom is unlikely to help them, preventing their integration process of science.

2.5. Benefits of interdisciplinary learning

Advocates of interdisciplinary learning (Field, Lee, & Field, 1994; Lattuca et al., 2004; Newell, 1994) have identified a wide range of educational benefits for students such as cognitive advancement and improvement of affective domains. First, interdisciplinary understanding facilitates higher-order thinking by mapping a variety of domains and asking students to notice meaningful patterns of information and ideas (Hursh, Haas, & Moore, 1983; Jacobs, 1989; Newell, 1994). According to the National Science Teachers Association (2003, p.18), high-order thinking is described as “the ability to engage in effective inquiry using scientifically defensible methods which is considered a hallmark of scientific literacy.” A growing body of research showed empirical evidence that interdisciplinary courses or programs benefit students to increase critical thinking (Astin, 1993; Buchbinder et al., 2005; Nowacek, 2005), meta-cognitive reflection (Wolfe & Haynes, 2003), problem-solving, and other higher order thinking skills (Boix Mansilla & Duraisingh, 2007; Lattuca et al., 2004; Leonard, 2007). Newell (1994) suggested that interdisciplinary learning accords with “enhanced affective and cognitive abilities, increased understanding of multiple perspectives, greater appreciation for ambiguity, and superior capacities for critical thinking” (p.35). Newell and Green (1998) also highlighted the fact that interdisciplinary learning leads to students’ deductive reasoning and reasoning by analogy and, in particular, synthetic thinking.

Second, interdisciplinary learning provides the additional richness of viewing the

topic through multiple lenses (Liu, Lee, Hofstetter, & Linn, 2008). Deep interdisciplinary understanding developed in a variety of distinct contexts taps cognitive processes, which in turn makes links between individual disciplines. The integration of disciplines helps students develop an essential core of knowledge for seeing the “big” picture of knowledge connections.

Third, educational scholars have suggested that interdisciplinary connections may make “learning easier . . . more realistic and potentially more useful to the student” (Shell et al., 2009, p. 184). In contrast, there was criticism of isolated disciplines as static and not reflecting the reality of every experience (Braunger & Hart-Landsberg, 1994; Hurd, 1991; Nielsen, 1989; Tanner, 1989). Interdisciplinary learning helps students make sense of scientific issues and problems presented in real-life contexts and aids them in coping with the issues by using skills and knowledge associated with any of the relevant disciplines.

Additionally, interdisciplinary education serves as one of the cornerstones of the move toward creating a learning environment where students are motivated to learn. A number of studies have been attentive to affective gains made in interdisciplinary learning contexts. Bragaw, Bragaw, and Smith (1995) confirmed the positive values of interdisciplinary learning on students’ attitudes and motivation in learning. Barab and Landa (1997) also stated that an interdisciplinary curriculum focusing on problem solving skills motivates students to learn. For example, students can be motivated by realizing that knowing a single discipline is insufficient for solving a given science problem, which allows them to actively take an interdisciplinary stance to develop their own central

perspective in science disciplines. Most science educators realized that if science lessons focus on the structure of the disciplines, students may have difficulty with taking interest or motivation, which eventually results in their poor academic achievement (Singh, Granville, & Dika, 2002; J. L. Smith, Deemer, Thoman, & Zazworsky, 2014).

An interdisciplinary approach in learning and teaching will also change the current assessment system. Discipline-based standardized tests may cause students to have difficulties with realizing that some science principles and concepts cross-cut with other science content areas. In contrast, interdisciplinary assessment breaks down strict disciplinary boundaries and makes students use multiple disciplines and integrate them to solve a question.

2.6. Definition of interdisciplinary understanding

Boix Mansilla et al. (2000) described the disciplinary understanding that individuals produce: “Individuals demonstrate disciplinary understanding when they use knowledge and modes of thinking in disciplines such as history, science, or the arts, to create products, solve problems, and offer explanations that echo the work of disciplinary experts.” (pp.17-18). Meanwhile, the idea of ‘interdisciplinary understanding’ is more elusive and difficult to define and there is little literature that mentions its definition and the cognitive processes involved in interdisciplinary understanding. Only Boix Mansilla and Duraisingh (2007) clearly proposed the following definition of interdisciplinary understanding:

The capacity to integrate knowledge and modes of thinking in two or more disciplines or established areas of expertise to produce a cognitive advancement—such as explaining a phenomenon, solving a problem, or creating a product—in

ways that would have been impossible or unlikely through single disciplinary means. (p. 219)

They believed that interdisciplinary understanding is a process that requires individuals to draw from two or more disciplines in order to solve problems, and offer explanations of the world around them in ways that would not have been possible through the scope of a single discipline. In contrast to the ‘multidisciplinary’ defined in the previous section, ‘interdisciplinary’ thinking requires more connected disciplines to provide some support for problem-solving. Based on the theoretical definition of interdisciplinary understanding described by Boix Mansilla and Duraisingh, I proposed an operational definition of the interdisciplinary understanding before creating an assessment (see chapter 3). Since interdisciplinary understanding is a latent construct that cannot be measured directly, it is necessary to obtain the operationalized definition, which refers to a specific measure that can easily be made observable. For example, the theoretical definition of ‘intelligence’ is an ability to learn or understand from experience and the operationalized definition can be the “score on the Stanford-Binet Intelligence test”.

2.7. Studies on interdisciplinary learning and teaching in science education

The objective of this section is to systematically identify and discuss empirical research on interdisciplinary learning, teaching, and assessment in science education. Before searching the literature, several inclusion criteria were formulated. First, each publication should be relevant to the literature that examines interdisciplinary learning, teaching, and assessment in secondary and higher education. Second, each publication should be peer-reviewed. Third, only publications written in English were included. Finally, the time span of the literature search was limited to 1990-2016 to provide an

overview of the most recent research in the field. In order to develop the search strategy appropriate to the main purpose of this review, various search terms were used after careful consideration. The following search terms were identified as being the most informative: “Science and (interdisciplinary or multidisciplinary or cross disciplinary or integrated) and (learning or teaching or thinking or understanding or education or assessment) and (middle school or junior high or high school or secondary or university or college or higher education or undergraduates)”. Quotation marks were used to search for the exact phrases. The chosen search terms were restricted to title, abstract, and keywords in order to obtain the literature with a clear focus on teaching, learning, and assessment within the context of interdisciplinarity. In the EBSCO database, sub-data bases were selected: the Educational Resources Information Center (ERIC), Education Source, PsycINFO, and Academic Search Complete. Additionally, the Web of Science, ProQuest for searching dissertations and theses, and Google Scholar were used. Finally, I reviewed recent conference papers on the study of the interdisciplinary learning in science.

According to the search results, the past decade has witnessed a surge in interdisciplinary research in higher education. However, there is scarce literature that represents empirical and theoretical evidence of interdisciplinary learning and teaching in the secondary education level. Moreover, literature on interdisciplinarity within science disciplines was rarely found in middle and high school levels. Only a handful of studies have looked at the nature of interdisciplinarity within science and other disciplines, especially engineering and mathematics. For example, Johnston, Riordain, and Walshe

(2014) placed significance on an integration approach between science and mathematics in secondary schools in Ireland. They contended that science and math are integrated in nature through mutual relationships. They also assumed that science can provide students with concrete examples for abstract mathematical concepts, while math can lead students to achieve a deeper understanding of science concepts by providing ways to quantify relationships among science concepts. Moreover, science activities containing math concepts can provide relevancy of related topics and promote motivation for learning math. Johnston et al. (2014) developed a unit of learning on distance, speed, and time and allowed three teachers to implement the unit in their classrooms. The study evaluated the teachers' perspective of the integration of math and science teaching through teachers' group interview and independent lesson observation. The key finding is that teachers thought the integration of science and mathematics in post-primary education facilitated authentic learning experiences for the students. This study confirmed the need and willingness of teachers to engage in continuous professional development to enhance their interdisciplinary experiences.

Munier and Merle (2009) presented an interdisciplinary mathematics–physics approach to the acquisition of the concept of angle by children in Grades 3–5. They hypothesized that physics-based situations help the children construct a geometry concept, angle in elementary math classes. For example, if the learning goal is to discover the law of reflection of light off a mirror, the children are able to grasp the equality of the angle of incidence and the angle of reflection. An implication of this study is to create connections between relevant subjects that help pupils to fully conceptualize

an existing concept in one of the subjects, in this study, angle.

Nagle (2013) emphasized an importance of preparing students for the interdisciplinary nature of modern biology and developed modules and courses for inquiry-based interdisciplinary learning in the Science Education for Public Understanding Program (SEPUP). This program focused on developing a secondary science curriculum and assessment with the interdisciplinary biology and environmental issues relevant to students' lives. Additionally, she proposed some elements that prevent the implementation of the interdisciplinary teaching in secondary science classrooms: teachers' and administrators' lack of perceptions regarding what students need for the next phase of their education, and teachers' lack of preparation to teach across disciplines. Nagle (2013, p. 146) suggested some recommendations for science teachers to foster their interdisciplinary teaching: 1) Develop rich examples related to overarching themes, problems, or socioscientific issues, 2) Encourage colleagues or policy makers to think towards interdisciplinary perspectives, 3) Emphasize the importance of evidence and logic in all sciences, 4) Promote the development of classroom and standardized assessments that go beyond memorization, 5) Support teachers to enact interdisciplinary curriculum, 6) Focus on developing a coherent K-12 science program to ensure a strong foundation at the school and system-wide levels. Furthermore, she recommended a concerted and long-term effort to change the current system of curriculum, instruction, assessment, and professional development.

Knapp, Desjardins, and Pleva (2003) designed a new chemistry curriculum for geology students. The curriculum was designed to use geological examples integrated

into basic chemical principles. As an example, students produced their own bronze alloy in the lab using copper carbonate, tin oxide, a flux (calcium carbonate) and a reducing agent (feed corn). During this activity, they were able to learn about the processes of smelting and write balanced redox equations and further determine the relative composition of copper and tin in their bronze. Knapp et al. (2003) indicated that the traditional chemistry courses mainly focus on the biological or molecular context rather than geological systems. This problem, too often, allows geology students not to make the connection between the context of the traditional chemistry courses and applications of same principles in Earth sciences. This study also implied that interdisciplinary endeavor on curriculum reform is essential to teach geoscience courses in college.

Rice and Neureither (2006) developed a semester science course (Science for Elementary Schools) where they wanted to integrate topics of Earth and space science (i.e., astronomy, meteorology, hydrology, geology) for pre-service K-8 teachers' integrated understanding of Earth science. They stated that this approach requires the basic concepts of physical science and life science as a foundation for understanding Earth system science. They did not want the course to be an exploration of discrete topics with no coherence or integration between science topics. The concepts of how matter changes, how energy flows, and how energy interacts with matter in physical and chemical changes were woven into every topic for the basis of weather, ecosystems, and human physiology as well as other science disciplines. For example, the concept of chemical changes of matter can be illustrated in the weather unit when discussing ozone depletion, acid rain formation, global warming, and ground level ozone. Similarly, energy

transfers and transformations can be illustrated in the weather unit regarding light energy being transformed into heat energy when it is absorbed by the surface of the Earth. The same physical science concepts appear again in geology, such as when acid precipitation falls on carbonate-containing rocks and heat is transferred in the mantle. At the end of the course, students were allowed to assess their own ability to integrate the various spheres of Earth system science in a project called the “biome” project. Each student chose a biome, from five biomes suggested (desert, prairie, deciduous forest, coniferous forest, tundra), that are prominent in the U.S. The biome project allowed each student to explain his or her biome in terms of the relevant basic concepts of astronomy (the sun’s path and intensity across the seasons), weather, hydrosphere (surface and ground water), geosphere (surface features and bedrock), plants’ and animals’ adaptations, and human activities and their impacts on the biome. To explain what would happen in each biome, this study assumed that the pre-service teachers should be equipped with relevant concepts from other science subjects that were aforementioned (e.g., utilization of matter and energy). Although this study did not explicitly mention that the project builds students’ interdisciplinary understanding, each project was used to improve the pre-service teachers’ interdisciplinary understanding.

Clary and Wandersee (2007) investigated effects of an integrative geobiological study in an introductory college geology course through pre-and post-instructional assessment in control and experimental classes. They had a hypothesis that students would show greater achievement in content knowledge in an integrated science course rather than a science course with traditional curriculum. Thus, this study attempted to

ascertain whether a science construct -petrified wood-, interconnecting geological and biological concepts, could facilitate students' understanding of fossilization, geologic time, and evolution theory, and eventually, earth's complexity. The Petrified Wood Survey was specifically designed to measure students' prior geobiological entry knowledge of fossilization, geologic time, and evolution. This study developed questions to ascertain their knowledge about the properties, formation, and chemical composition of petrified wood. The experimental class group who received integrated petrified wood instruction revealed statistically significant knowledge gains about petrified wood's abundance, properties, nature, location, and geologic time except the understanding of fossilization. This study showed that interdisciplinary programs improve students' achievement scores in comparison to a traditional disciplinary lesson. Additionally, McComas and Wang (1998) informed the benefits of the blended (i.e., integrated) science teaching from a number of perspectives including philosophical, psychological, pedagogical, and pragmatic justifications that are beyond students' academic achievement. They highlighted that students' developmental journey and processes of blended science learning result in epistemological benefits by making connections within and across disciplines and seeing the bigger picture.

Taber (2005) proposed the definition of 'conceptual integration' (CI) as "the knowledge structures of an individual organized in such a way that there is strong linking between different areas". Literally, the perspective of CI is on a par with that of interdisciplinary understanding. He investigated the extent to which students achieve conceptual integration of the science (Taber, 2008). His work was based on the

framework of constructivism models for learning to support the process of conceptual integration, supporting complex systems of nature. He used an interview protocol to explore the level of integration in the students' science knowledge across college-level subjects (chemistry and physics). This study emphasized that students should not only learn individual scientific principles and ideas, but also connect knowledge structures across other disciplines conceptually.

In addition to Taber's studies, some literature has focused on science teachers' CI. For example, Ganaras, Dumon, and Larcher (2008) examined the mastering of the chemical equilibrium concept by prospective physical science teachers. The 'chemical equilibrium' is an integrating and unifying concept in science because it requires the connection of several concepts concerning varied domains of chemistry (e.g., thermodynamics, kinetic, structure of matter, etc.). Tuysuz, Bektas, and Geban (2014) examined how pre-service science teachers think about CI. The study showed that all pre-service teachers could not even conceptualize the meaning of conceptual integration and did not prefer to use CI for teaching because it was not stated in the curriculum.

2.8. Literature on assessments associated with interdisciplinary understanding in science education

During the past decades, there has been a growing emphasis on interdisciplinary learning and teaching at both secondary- and college-level education, yet few empirical research studies of interdisciplinary assessment were conducted. One scholar even described assessment as "the 'black hole' of interdisciplinary education" (Boix Mansilla, 2005, p. 18). Through a rigorous search, only one research study was found that dealt

with interdisciplinary assessment in measuring interdisciplinary understanding of a higher education context and it was recently published in a science education journal (Shen, Liu, & Sung, 2014). Shen et al. (2014) also argued that there has been a lack of empirical research assessing students' interdisciplinary understanding in areas of science education and emphasized the necessity of assessments for interdisciplinary understanding. The authors developed an interdisciplinary assessment that targets college-level science and assessed students' interdisciplinary understanding of osmosis, which involves knowledge from multiple science disciplines. They used the theoretical framework of knowledge integration for interdisciplinary understanding including three special processes: translation, transfer, and transformation of knowledge. The assessment outcome and its analyses obtained by a Rasch partial credit model (PCM) showed that the items demonstrated satisfactory psychometric properties and revealed the differences between students' disciplinary and interdisciplinary understanding. The findings demonstrated that interdisciplinary understanding of the college students was very limited compared to the growth of their disciplinary knowledge. This study has some differences in that the "knowledge integration (KI)" as the conceptual framework for interdisciplinary understanding and the assessment was implemented as homework assignments.

The small body of literature of interdisciplinary assessment in science areas allowed me to expand the search to assessments with respect to 'knowledge integration (KI)' and 'conceptual integration (CI)'. Liu et al. (2008) developed a KI assessment asking students about complex thinking involving science inquiry skills such as linking,

distinguishing, evaluating, and organizing their ideas. The assessment items readily contain contents that are taught in six common science courses (physics, life, Earth sciences for middle school students, and physics, biology, and chemistry for high school students). Items were selected in existing standardized tests such as NAEP and TIMSS, and previous knowledge integration research. Psychometric properties of the items were analyzed, using the Rasch partial credit model.

Lee and Liu (2010) assessed the development of student understanding, using a knowledge integration construct, ‘energy’. This study adapted physical, life, and Earth science contexts in middle school grades and built items addressing energy source, transformation, and conservation from published standardized tests. The analyses by a Rasch partial credit model indicated that conservation items are associated with the highest knowledge integration levels, followed by transformation and source items. The origin of the item difficulty is in part related to the increase of integrated components. This means that when students develop concepts about energy sources and transformation processes, they can learn energy conservation in a more integrated manner by recognizing various energy sources and associating changes of the system. In addition, Lee and Liu (2010) revealed the overall level of knowledge integration of middle school students and the difference in the knowledge integration across grade levels. However, although they adopted the multi-science disciplines, the items do not seem to have interdisciplinary features. Liu et al. (2015) also reported the development and validation process to measure integrated understanding of energy while implementing an inquiry-based curriculum, using a two-year longitudinal data. 6th to 8th grade science teachers

administered the beginning-of-year assessment to evaluate students' prior knowledge of energy. After implementing one or more of the energy lesson units, the teachers administered the end-of-year assessment to measure the enhanced student learning. The units were designed to promote the students' ability to make connections among energy sources, energy transformation, and energy transfer across science topics. The psychometric properties of the items were acceptable. For the longitudinal cohort, both 6th and 7th graders made significant progress from Year 1 and Year 2.

Schaal, Bogner, and Girwidz (2010) monitored 9th graders' abilities to construct integrated knowledge of mammalian hibernation strategies with both biological and physical perspectives. They indicated that science educators in their classrooms can easily assess students' interconnected concepts across science subjects by using computer-assisted concept maps. An individual's knowledge structure may be regarded as a concept map, which represents the aspects of an individual's integrating thinking. Their work showed some potential of promoting interdisciplinary abilities of learners through interdisciplinary instruction.

2.9. National and international tests and project-based assessments

The Framework (2012) and the NGSS (2013) guided educators significantly in rethinking interdisciplinary learning in science education. This perspective may lead to national, state, and even international assessments to explore new approaches. In order to see whether current and ongoing assessments have interdisciplinary aspects that meet the standards, I reviewed recent national and international tests first, and then the existing assessments for project-based learning with interdisciplinary nature. The National

Assessment of Educational Progress (NAEP) science assessment is the largest nationally representative assessment that monitors the overall progress of science learning and teaching in U.S. classrooms. A new type of assessment in 2009 was administered through the usage of computer interactive and hands-on tasks. The task formats are closer to what is required for measuring performance expectations in the NGSS (NRC, 2014), but NAEP items assessed inquiries separately from science content rather than an integrated understanding across diverse scientific fields (Liu et al., 2008).

The Program for International Student Assessment (PISA) aims to measure scientific literacy of students: an individual's scientific knowledge (knowledge of science and knowledge about science) and the use of that knowledge to explain scientific phenomena and to draw evidence-based conclusions about scientific issues (OECD, 2013). The assessed knowledge was selected from major science disciplines such as physics, biology, chemistry, and Earth and space science and real-life contexts such as 'acid rain' or 'greenhouse'; however, almost all the items were created from one discipline. For example, the 'acid rain' item assessed students' knowledge in "physical science" and 'greenhouse effect' was only concerned with knowledge of Earth and space systems. This implies that the science items in PISA are not aligned with an interdisciplinary perspective.

The Trends in International Mathematics and Science Study (TIMSS) science assessment is organized around two domains: (1) content knowledge about the subject matter and (2) cognitive or thinking processes. At grade 4, TIMSS assesses student knowledge in three content domains: life science, physical science, and Earth science and

at grade 8, it assesses student knowledge in four content domains: biology, chemistry, physics, and Earth science. As shown in Table 1, a proportion of item score points is attributed to each content domain, which implies the TIMSS framework is unlikely to capture attainment of interdisciplinary learning that emphasizes connections among scientific concepts.

Table 1. Percentage of TIMSS science assessment score points at grade 4 and 8 devoted to content and cognitive domains in 2011.

| Grade 4 | | Grade 8 | |
|------------------|-----------------------|-----------------|-----------------------|
| Content domains | Percent of assessment | Content domains | Percent of assessment |
| Life science | 45 | Biology | 37 |
| Physical science | 35 | Chemistry | 20 |
| Earth science | 21 | Physics | 25 |
| | | Earth science | 18 |

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

I also examined recent assessment research to figure out the trends in measuring interdisciplinary understanding: curriculum-embedded assessments from the Investigating and Questioning our World through Science and Technology (IQWST) project by Krajcik, Reiser, Fortus, and Sutherland (2013); the Principled Assessment Design for Inquiry (PADI) system; a project at the Berkeley Evaluation and Assessment Research (BEAR) Center, and Technology Enhanced Learning in Science (TELS). IQWST units were designed to assess middle school students in investigating phenomena and engaging in scientific argumentation, and building models as they explore disciplinary core ideas in depth. All curriculum units of IQWST were not designed to be able to support several connections across disciplines, but some units allow students to draw on interdisciplinary ideas from different science disciplines. For example, if

students are learning about how food is used by organisms in the context of the life sciences, they should have the relevant ideas about chemical reactions (e.g., release of energy) in the context of physical science such as conservation of matter and transformation of energy. Explanation items especially require building knowledge that links core ideas across several science disciplines to develop and defend students' explanations based on evidence (NRC, 2014).

An additional practical and theory-based assessment of science inquiry is the PADI guided by Haertel and Mislevy (2006). The PADI assessment was created through the systematic design to measure scientific inquiry. The PADI project is based on cognitive psychology research on scientific inquiry and measurement theory to formulate a structure of inquiry assessments. The PADI includes three essential components: student models, evidence models, and task models. The development of the PADI was based on the Evidence Centered Design framework (Mislevy, Almond, & Lukas, 2003). More specifically, the student model defines the latent traits or constructs to be measured. The evidence model determines the performance level of the student on the constructs defined in the student model. Finally, the task model specifies procedures for the development of an assessment that is aligned with the student and the evidence models. Therefore, the Evidence Centered Design framework provides (1) consistency with the developer's goals/intentions and (2) internally coherent guidelines for constructing assessments and interpreting them (Baxter & Mislevy, 2004). The evidence-centered design approach looks at the interaction between content and skills in order to discern, for example, how students reason about a particular content area. Ideally, this approach

yields test scores that are very easy to understand because the evidentiary argument is based not on a general claim that the student “knows the content,” but on a comprehensive set of claims that indicate specifically what the student can do within the given domain. Claims that are developed through this approach can be guided by the purpose for assessment (e.g., to evaluate a student’s progress during a unit of instruction, to evaluate a student’s level of achievement at the end of a course).

While the PADI used the evidence-centered design, the Berkeley Evaluation and Assessment Research (BEAR) Assessment System applied construct modeling. Construct modeling contains four building blocks that start with a construct map. The construct map is used to guide the design of assessment items and to describe the developmental stages of scientific understanding aligned with curriculum goals. The item design considers multiple types of assessments that can effectively elicit students’ knowledge as well as a coding scheme for their responses. The outcome space is where the graders decide how to draw inferences from the “raw” responses, and how the responses can be scored using the scoring rubric, which is designed to ensure that students’ responses can be interpreted in light of the construct map. The final building block of the BEAR Assessment System is the measurement model. In this step, the BEAR system used Rasch modeling to show psychometric evidence of the validity and reliability of the assessment (Wilson, 2005).

Lastly, the Technology Enhanced Learning in Science (TELS) assessments ask students to connect scientific ideas and give explanations in varied contexts, which is consistent with knowledge integration framework. The explanation items of TELS tests allow students to explain natural phenomena and to elicit scientific inquiry and reasoning

from linking everyday experiences. For example, the blown fuse context connects to students' everyday experience. TELS Director Marcia C. Linn and her colleagues characterized Knowledge Integration (KI) as the basis for TELS curricular projects and assessments. The KI items provide opportunities for students to link their ideas about scientific phenomena to assess more coherent explanations of the ideas; however, the items are not required knowledge from multiple science disciplines.

This section investigated national, international, and project-based assessments with interdisciplinary aspects. The aforementioned assessments except for the TELS assessment consist of totally discipline-based items in each science discipline, showing weak connections among other science areas.

2.10. Assessment triangle

Assessment, as defined in the National Science Education Standards (NRC, 1996), is “a systematic, multi-step process involving the collection and interpretation of educational data” (p. 76). Assessment specialists believe assessment as a process of reasoning from evidence—“of using a representative performance or set of performances to make inferences about a wider set of skills or knowledge” (NRC, 2014, p. 48). The National Research Council (NRC, 2001) portrayed this process of reasoning from evidence as a triangle with three corners—cognition, observation, and interpretation to emphasize their connected relationships (see Figure 3). Cognition, in an assessment design, is “a theory or set of beliefs about how students represent knowledge and develop competence in a subject domain” (NRC, 2001, p. 44), which are important to measure. In measurement terminology, the assessed knowledge and skill is referred to as

“constructs”. An assessment should start from an explicit and clearly well-defined construct because the design and selection of the tasks need to be tightly linked to the specific inferences about student learning. If the intended constructs are clearly specified, the design of specific items or tasks and their scoring rubric could provide clear inferences about the students’ capabilities.

A second corner of the triangle is the observation of the students’ capabilities in a set of assessment tasks designed to show what they know and can do. They are based on theories and beliefs concerning knowledge and cognitive processes to acquire valid and rich responses. Thus, observations support the inferences that will be made based on the assessment results. The *Interpretation* vertex includes all the methods and tools to infer the results of observations that have been collected. Statistical models or qualitative models can be used for methods or tools to interpret the patterns of the data collected through assessment tasks. The interpretation model needs to fit the type of data collected through observation. Through interpretation, the observations of students’ performances are synthesized into inferences about their knowledge, skills and other attributes being assessed. The method used for a large-scale standardized test might involve a statistical model. For a classroom assessment, it could be a less formal method of drawing conclusions about a student’s understanding on the basis of the teacher’s experiences with the student, or it could provide an interpretive framework to help make sense of different patterns in a student’s contributions to practicing and responding to questions. Pellegrino (2012) asserted in the NRC report *Knowing What Students Know*: “These three elements—cognition, observation, and interpretation—must be explicitly connected and

designed as a coordinated whole. If not, the meaningfulness of inferences drawn from the assessment will be compromised” (p. 2). It is recommended that assessments should be equipped with a design process that coordinates the three elements of the triangle, not only focusing on the observation vertex, to support the intended inferences.

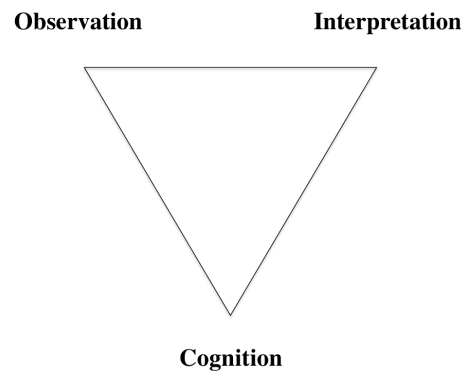


Figure 3. Assessment triangle.

2.11. Construct-modeling

Kind (2013) argued that science assessment should change from an item-driven practice to a construct-driven one. The assessment triangle provides a powerful approach to the nature of assessment, but it is not satisfactory for creating an assessment tool towards the construct-driven practice (Kind, 2013). Wilson and his colleagues (2005) provided a framework for developing assessment tasks that take into account the evidentiary reasoning logic: construct-modeling, emphasizing the importance of constructs. Construct-modeling contains four building blocks for guiding a assessment design process: construct map, item design, outcome space, and measurement model (Wilson, 2005). The construct map embodies the first of the four principles: that of a

developmental perspective on the assessment of student achievement and growth. A construct “can be part of a theoretical model of a person’s cognition ... their understanding of a certain set of concepts” (Wilson, 2005, p. 6). The constructs that capture deep understanding in science can be defined as what is measured or assessed (Wilson, 2008). Just as in learning progressions (NRC, 2007), the construct map is a well-thought-out ordering of qualitatively different levels of performance in a domain. A distinct difference between building blocks of the construct modeling and the assessment triangle is that the building blocks emphasize that assessments need to be based on a developmental perspective on student learning. The inception of construct maps is a combination of research in cognitive structure and adept judgment on the varying levels of performance or competence, which are further solidified by empirical research on students’ responses and performance. The second building block is the item design, which contains a description of the possible forms of items that will be used to elicit evidence about students’ knowledge and understanding. The third building block is the outcome space, a description of the qualitatively different levels of responses to items and tasks that are associated with different levels of the construct. The outcome space shows the process of developing scoring rubrics including criteria for judgment of students’ response and interpretation of the items (Wilson, 2005). The last building block is the measurement model. In this step, statistical models are used to relate the scores earned on items and tasks back to the construct map. The four building blocks become part of a development cycle in the assessment task, where with continuous successions the cohesion among the blocks is getting better and clearer (see Figure 4).

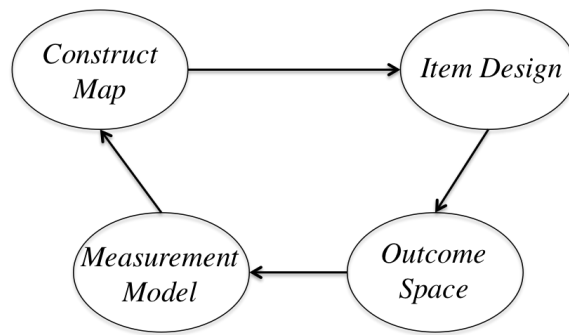


Figure 4. Four building blocks of construct modeling (Wilson, 2005).

The construct modeling approach is called the “principled” approach to assessment design in that it provides a systematic approach to designing assessment tasks (NRC, 2014). Using this approach for designing a construct-centered assessment task can specify a construct of interest, develop a construct map, and support better inferences in the interpretation measuring a student’s proficiency. Through the rigorous development process of an assessment, construct validity of the assessment tool can be established.

2.12. Validity

Validation is the process of evidence-based judgment that supports the appropriateness of the inferences that are made of student responses for specified assessment uses. Validity refers to the degree to which the evidence supports that these interpretations are correct and that the manner in which the interpretations are used is appropriate (American Educational Research Association [AERA], American Psychological Association [APA], & National Council for Measurement in Education [NCME], 2014). The validity perspective in the *Standards* (AERA et al., 2014) discussed

five sources of validity (based in part on Messick, 1989): evidence based on (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing.

Messick (1989) defined construct validity as: “the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and relationships with other variables” (p.34). The traditional concept of validity focuses on “types of validity”, which have been grouped into three categories; content-related, criterion-related, and construct-related (Messick, 1995). However, Messick (1995) argued that the traditional concept of validity is fragmented and incomplete, suggesting a unified nature of validity: construct validity. He defined construct validity as “an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment” (p. 741). Messick (1995) provided six facets of construct validity: content, substantive, structural, generalizability, external, and consequential. This study uses Messick’s (1995, 1996) framework for validity, six aspects of construct validation to guide the validation of the interdisciplinary assessment in carbon cycling.

- Content validity: According to AERA et al. (2014), evidence based on content is “obtained from an analysis of the relationship between a test’s content and the construct it is intended to measure” (p. 11). Messick (1989, 1995) defined the content aspect of construct validity as evidence of content relevance, representativeness, and technical quality. Validity evidence of the content

typically was done through an analysis of the target content, the curriculum, and competencies actually present in assessment items (Messick, 1996). Particularly, content domains should be relevant and representative of content standards.

Experts perform this judgment of construct/content validation. For example, the carbon cycling assessment typically should be based on the NGSS that specifies what it is that student should learn and teachers should teach.

- Substantive validity: In order to obtain substantive validity evidence, test developers must first consider what response processes and skills are necessary to complete tasks within the construct. Response processes refer to “the procedures, strategies, and cognitive behaviors that an examinee engages in while responding to a test item” (Lai, Wei, Hall, & Fulkerson, 2012, p.10). As an assessment task needs to be representative and relevant to content specifications, so intended processes used in completing the tasks need to be representative and relevant to the processes that constitute the construct of interest (Miller & Linn, 2000). For instance, if a test developer believes that an interdisciplinary thinking process is needed to solve a construct-response item, the test developer may infer that the student who answers the item correctly uses interdisciplinary thinking processes to solve the problem. One way to gather substantive validity evidence is to have students participate in “think aloud” interviews, in which they verbally communicate their thinking processes as they complete assessment tasks (Messick, 1995; Nitko & Brookhart, 2010).

- **Structural validity:** According to AERA et al. (2014), “analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based” (p. 13). Internal structure evidence focuses on how well the items correlate with one another (Messick, 1995). Items that show a strong correlation suggest that they are measuring the same thing (e.g., knowledge, response processes, etc.); on the other hand those with a weak correlation suggest that they measure different things. The overall goal for an assessment tool is that all of the items are to independently measure one dimension or construct (Nitko & Brookhart, 2010). In addition to unidimensionality, strong rubric structure with high levels of rater consistency shows structural validity evidence.
- **Generalizability:** When creating an assessment, researchers must generalize inferences derived from the assessment scores. Different demographic information or contextual factors (e.g., a low socioeconomic status sample versus a high socioeconomic status sample, different raters) lead to different inferences on the assessment (Kane, 2006; Nitko & Brookhart, 2010). Differential item functioning (DIF) analysis is concerned with identifying significant differences across subgroups (e.g., commonly gender or ethnicity). The DIF provides helpful evidence to determine the measurement bias across the groups, identifying assessment items that are differentially difficult for examinees who have the same ability in regard to a construct.

- External validity: The external aspects of validity refer to the extent to which assessment scores are related to other external measures (AERA et al., 2014). Messick (1996, p.11) argued that the constructs represented in the assessment “should rationally account for the external pattern of correlations” (Messick, 1996, p. 11). Convergent evidence is based on correlations between the test scores and the same or similar constructs of other measures. Discriminant evidence is based on correlations between test scores and measures of different constructs (AERA et al., 2014).
- Consequential Validity: The consequential aspect of validity includes evidence and rationales for evaluating the intended and unintended consequences of test use and their impact on score interpretation (Miller & Linn, 2000). Intended consequences might include the ease of use, instructional benefits, and consequences for students, etc. (Linn & Baker, 1996). Unintended consequences might include bias in the assessment, leading to misinterpretations for some subpopulation (Bond, 1995). Popham (1999) argued that “consequences should be systematically addressed by those who develop and utilize tests, but not as an aspect of validity” (p. 9). Despite this argument, other scholars argued that the consequential aspects of validity can help indicate how, when, and where it should appropriately be used (Haladyna, 2006; Nitko & Brookhart, 2010). For example, there is a hypothesis that the use of the first-term introductory biology exam offers meaningful information to permit student enrollment in second-semester biology. If an original logistic regression model in the study shows that

the student performance of the first-term biology exam did relate to their likelihood of passing the second-term biology exam, this result supports the consequential validity of the first-term exam.

Chapter 3: Methods

This chapter focuses on strategies used to design and implement assessment tasks that measure the intended interdisciplinary understandings of science based on the construct-modeling framework. Section 3.1 describes the overall research design. Section 3.2 describes the sample and data collection process. For methodological and systematic coherence in designing the assessment tasks, Wilson's Construct Modeling approach (2005) was used as an assessment model to align cognition of interdisciplinary understanding (section 3.3) with item design (section 3.4) and scoring rubric development and validation (section 3.5). Section 3.6 reports steps of data analyses for interpretation of student performance data using a measurement model.

3.1. Overview of the research design

In this section, I detail the systematic process used to develop and validate the interdisciplinary science assessment (ISA) instrument. Multiple procedures were carried out within each stage. Figure 5 provides a graphic overview of this progression. The subsections below provided explanations of how the construct map, content experts, test administration, scoring rubrics, confirmatory factor analysis and item response theory were used to develop the instrument, and to establish validity and reliability.

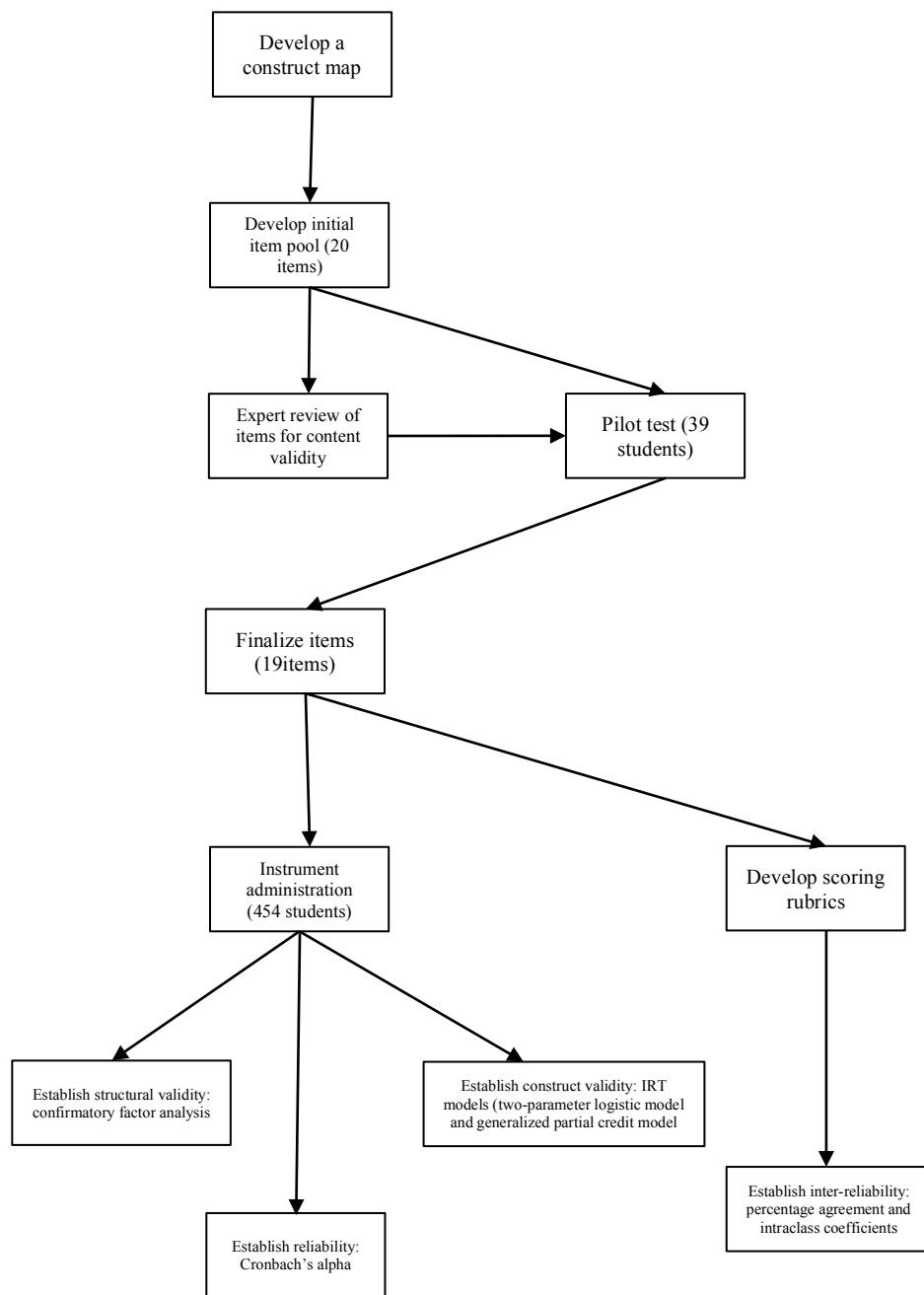


Figure 5. Graphic display of the instrument development and validation process.

3.2. Participants and data collection

The assessment data from 44 high school and 410 college (including four graduate) students were collected in a public high school and a public research university in Texas during the 2015-2016 school year. Participants initially read an informed consent form and then took the test survey. All test responses were collected by a web-based Qualtrics system. Demographic data for the sample are listed in Table 2. The 454 students showed variation in their demographic information including gender, race, and grade level. These participants ranged in grade levels from 9th grade high school students to graduate students. The reason for targeting high school students and college students is based on the assumption that they have enough prior disciplinary knowledge for interdisciplinary understanding, allowing for the possibility that successful interdisciplinary understanding of science must be built upon the success of disciplinary science. Of these students, 41.9% were males and 58.1% were females. The racial diversity of the participants was: White (39.0%), Asian (28.9%), Hispanic or Latino (23.3%), African American (5.1%), Native Hawaiian or other Pacific Islander (0.4%), and other (3.3%).

Table 2. Demographic information.

| | N | Percent |
|--------------------------|-----|---------|
| Grade level | | |
| High school | 44 | 9.6 |
| College/Graduate | 410 | 90.4 |
| Gender | | |
| Female | 264 | 58.1 |
| Male | 190 | 41.9 |
| Race | | |
| White | 177 | 39.0 |
| American Indian or | 0 | 0 |
| Asian | 131 | 28.9 |
| Native Hawaiian or Other | 2 | 0.4 |
| Black or African | 23 | 5.1 |
| Hispanic | 106 | 23.3 |
| Other | 15 | 3.3 |
| Overall | 454 | 100 |

3.3. Interdisciplinary understanding construct map

It is important to establish an operationalized definition of interdisciplinary understanding to develop a construct map at the initial stage of assessment development. This study thus started off by describing the operationalized definition of ‘interdisciplinary understanding’ in this study. This definition was created based on the theoretical definition provided by Boix Mansilla and Durasing (2007):

The performance of high school and college students to use multiple sub-science disciplines and integrate knowledge from different sub-science disciplines to explain a scientific phenomenon in the context of carbon cycling, or to develop an argument about a scientific problem that cannot be dealt with adequately by a single discipline.

A construct map based on the operationalized definition should elaborate the correspondence between each student response and a certain growth level. In this study, the construct map shows descriptions of the four hierarchical levels regarding

interdisciplinary understanding and example responses of the students: more interdisciplinary concepts at the top of the map and more unidisciplinary-oriented ones below. The lowest level of the construct map, labeled “no response or irrelevant,” indicates instances when individuals leave their answer as blank or have only irrelevant disciplinary knowledge in a given scientific phenomenon. The next level of “unidisciplinary” captures students’ correct reasoning in only independent science discipline. At the “partially interdisciplinary” level, students use only partial science disciplines, not fully elaborating how the ideas are integrated between other science disciplines. The highest level is indicated as “fully interdisciplinary”, describing the level at which students are able to explain their scientific ideas with integration of all the necessary science disciplines. The specified levels of the construct map were developed on the basis of the combined knowledge gained from a KI framework originally proposed by Liu et al. (2008) and an interdisciplinary framework developed by Golding (2009).

3.4. Item design

Developing an instrument is a multi-step process requiring careful decision-making at each step. A five-step procedure used in developing the ISA is described in this section.

1) Creation of an initial item pool: The initial item pool for the interdisciplinary assessment was created based on following three tenets. First, a pool of potential items should be aligned with performance expectation on disciplinary core ideas (DCIs) shown in the NGSS. As the construct of interest in the assessment is ‘interdisciplinary understanding’ in science, the second tenet is that assessment items should contain either

unidisciplinary or interdisciplinary components of science to reveal a wide range of levels in students' interdisciplinary understanding. The third tenet involves items to address real-life problems, such as global warming and ocean acidification, that could help students demonstrate their scientific literacy through inquiries on real-life issues.

Additionally, it is necessary to figure out what core concepts are in the carbon cycling content domain to develop an item pool for the ISA. This study used 'concept maps' from content experts to reveal what they considered to be core concepts within carbon cycling and the integrated structures of the concepts. Vanides et al. (2005) stated that using a concept map in planning to find subthemes for assessment helps in making the assessment "conceptually transparent" to researchers. In this study, a total of seven faculty members from diverse natural science areas and two doctoral students, having a background in biology and physics, respectively, aided in the creation of concept maps by providing their expertise.

2) Establishing content validity: In order to ascertain the quality of items, it is necessary to establish content validity. Although there are numerous methods leading to agreement among experts regarding content validity, the most widely reported measure of content validity is content validity index (CVI), which quantifies experts' degree of agreement (Lynn, 1986). The CVI represents the proportion of experts who agree concerning content validity. Lynn (1986) recommended a four-point scale since it does not contain a neutral or ambivalent midpoint (i.e., 1= not relevant; 2= unable to assess relevance without item revision or item is in need of so much revision that it would no longer be relevant; 3=relevant but needs minor alteration; 4=very relevant). For each item, the CVI

is computed by the number of experts giving a rating of either 3 or 4 divided by the total number of experts. For example, an item that was rated as '3 or 4' by four out of five judges would have a CVI of .80. There are minimum criteria for the CVI proposed by Lynn (1986) (see Table 3).

The same nine experts reviewed the items of the initial version and participated in a follow-up interview. After carefully reading questions, the three tenets for the item development process, and the item characteristics, the experts were asked to rate the extent to which they agreed that each item measures carbon cycling content on the four-point rating scale. According to the minimum criteria of the CVI, seven out of nine experts needed to choose '3 or 4'. If an item did not meet the criteria, in other words, if the item did not achieve the required agreement from the experts, the item was eliminated or further revised for the final version of the ISA. In the interviews, the reviewers were asked about their overall impression of the items and their opinion about the alignment between the items and the carbon cycling concepts. They were also asked to offer modifications for the items that they felt needed improvement.

Table 3. Proportion of experts (above the line) whose endorsement is required to establish content validity beyond the .05 level of significance (Lynn, 1986).

| Number of experts | Number of experts endorsing item as content valid | | | | | | | | |
|-------------------|---|------|------|------|------|------|------|------|------|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2 | 1.00 | | | | | | | | |
| 3 | 0.67 | 1.00 | | | | | | | |
| 4 | 0.50 | 0.75 | 1.00 | | | | | | |
| 5 | 0.40 | 0.60 | 0.80 | 1.00 | | | | | |
| 6 | 0.33 | 0.50 | 0.67 | 0.83 | 1.00 | | | | |
| 7 | 0.29 | 0.43 | 0.57 | 0.71 | 0.86 | 1.00 | | | |
| 8 | 0.25 | 0.38 | 0.50 | 0.63 | 0.75 | 0.88 | 1.00 | | |
| 9 | 0.22 | 0.33 | 0.44 | 0.56 | 0.67 | 0.78 | 0.89 | 1.00 | |
| 10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 |

3) Pilot test: After obtaining evidence of content validity of the assessment, the second revised version of the ISA was developed. The total number of the items is 20 including 12 MC items and 8 CR items. It was piloted with 22 middle school and 17 high school students who participated in summer programs at a public research university and an SAT class held at a Presbyterian church in Texas. This pilot administration included 20 questions. The three main goals of the pilot test were (1) to obtain information about the meanings students made of the assessment items, (2) to identify the terms in the items that were not well understood or that could lead to response error, and (3) to find out how long it would take to complete the assessment in real time.

4) Reviewing and refining the item pool: As a result of the pilot test, some items were rewritten to make them clearer for students to understand, and one item asking “where most of Earth’s carbon resides” was removed, as this question required only recalling

facts and, in the pilot test, almost all students answered this question incorrectly. Three items were removed and two new items were added with content validity confirmation from the experts. In the final version, the number of items was 19 in total. According to Comrey (1988), in the case of a simple construct that is one dimensional in nature, 10 or fewer items will likely be enough to sufficiently measure the construct.

5) Administration procedure: During the administration process of the final items, assessment data from 44 high school students (grade 9th-12th) and 410 college/graduate students were collected during the 2015-2016 school year. I contacted several instructors who were teaching courses at a public university in Texas, and high school science teachers in the same region, to receive permission and support to recruit college and high school students. University students were recruited by an announcement or email from instructional staff. Three instructors from the biological sciences department and two instructors from the geoscience department helped in recruiting students in their courses. The students who gave their consent to the research participation were given a web-based test. The time limit for the test completion was 45-50 minutes based on the average time in which students completed the previous pilot test. The interdisciplinary assessment was administered, using the Qualtrics tool.

(https://utexas.qualtrics.com/SE/?SID=SV_5aPf8JhAfkCJcVr). The assessment in Qualtrics consisted of three parts: 1) the actual test, 2) a section asking about the students' background such as demographic information (gender, grade etc.) and the kinds of science courses they had taken, and 3) a consent form that informed them of the nature of the assessment.

There are some benefits for web-based tests rather than traditional paper and pencil ones. Integration of educational technology into assessment enhances the capacity of student Information and Communication Technologies (ICTs) literacy. For instance, as a practical reason, because many standardized tests are now offered electronically, using the web-based test format provides an opportunity to get used to the technology and web environment for students. Also, taking a web-based test is flexible in terms of time and place, because students can access it at their convenience if they have personal digital apparatus (e.g., laptop, smart phone, etc.) (Aggarwal, 2003).

3.5. Outcome space (rubric development and validation process)

Outcome space describes how individual test items are to be scored and how item scores are to be combined to yield overall test scores. For MC items, one of the response options is considered the correct response. CR items require a more clearly defined and fully elaborated scoring rubric to evaluate student responses. The scoring rubric represents a set of qualitatively different categories of student responses elicited by the items (Wilson, 2005), where the number of score levels employed and criteria for each score level are described. The rubric should be constructed with overall clarity through an iterative process. Note that there is some possibility to confuse the outcome space with the construct map. The fundamental distinction is that the construct map is defined at a more general level compared to the specificity of the scoring rubrics (Wilson, 2008).

Separate rubrics for each item were developed, using the following seven-step process. All seven steps comprise processes to ensure the reliability and validity of the rubrics.

1) Identifying the construct: The construction of rubrics should be based on clarity of what should be assessed. For doing this, the construct map developed in the previous stage was used as a basis to build a holistic scoring rubric (see Table 4). The holistic rubric addressed how many performance levels should be specified and what characteristics should be defined for each performance level. The construct map included the following descriptive categories: “no response or irrelevant”, “uni-disciplinary”, “partial,” and “full”, whereas the holistic rubric categorized a student’s interdisciplinary understanding as “non-disciplinary”, “uni-disciplinary,” or “interdisciplinary”. If the students’ answers are not correct or irrelevant or they write the prompt as it is, it can be assumed that these students do not have any interdisciplinary understanding on this level. Only partially correct or fully correct answers can be judged in regards to the degree of interdisciplinary understanding the students have. The levels of interdisciplinarity in the holistic rubric were more simplified than those shown in the construct map, collapsing “partially interdisciplinary” and “fully interdisciplinary” category into “interdisciplinary”. This process allows raters to avoid subjective decisions between two boundaries, leading to consistent scores. Higher scores on the rubric represent more interdisciplinary understanding of a scientific phenomenon based on relevant ideas from diverse but necessary science disciplines.

Table 4. Holistic rubric on interdisciplinary understanding.

| Correctness | Interdisciplinarity | Description |
|---|---------------------|---|
| Partially/Fully correct | Interdisciplinary | A student uses relevant scientific concepts and principles to explain a specific event in carbon cycling to demonstrate his/her reasoned interdisciplinary understanding. |
| | Uni-disciplinary | A student response has a partially or fully accurate understanding of the scientific concepts and principles from one necessary discipline to explain a specific event in carbon cycling. |
| Incorrect/ Off topic/Blank/ Restatement of the prompt | Non-disciplinary | A student response is incorrect/not relevant/blank. |

2) Identifying levels of performance and assigning scores: The elements in the holistic scoring rubric were translated into separate analytic rubrics. The analytic rubrics were presented in a way that allowed raters to objectively determine the level of interdisciplinary understating in students' responses. The determination of the perfect score for the items depends on to what extent the items have an interdisciplinary nature according to the construct map developed in the previous stage. The rubrics consist of two different score ranges, 0-6 points for two disciplinary items and 0-8 points for six interdisciplinary items, which denotes a progression from the absence of the interdisciplinary understanding to extensive use of knowledge in different science disciplines.

3) Writing a description for each point on the rubric scale: The initial analytic rubrics were drafted based on the scientific content and analysis of student artifacts. In constructing the provisional rubrics, the highest and lowest level descriptions were decided before the middle-range ones for each item were added. Further, based on a constant comparison approach (Glaser & Strauss, 1967), I randomly selected a sample of

50 responses for each item, and carried out comparisons among the students' responses and sequenced their levels depending on their response patterns. Some examples of student responses were also included at each achievement level.

4) Obtaining feedback on the rubric and carrying out revisions: Five experts from Earth science, biology, physics, and chemistry participated in reviewing the initially developed rubric. The experts were asked to comment on the clarity of descriptions of ideal answers and appropriateness of the rating levels. For example, a suggestion that came from an instructor with regard to an incorrect scientific fact was to replace the sentence such as “it helps the Earth hold on to more infrared (IR) radiation from the sun, which in turn warms the climate further”, to “it helps the Earth hold on to more infrared (IR) radiation reflected back to the Earth by the atmosphere, which in turn warms the climate further”, indicating a possibility for improvement. The three experts made suggestions that step II also could show the decomposition or fermentation processes beyond the process of cellular respiration in item 2; thus, I added descriptions of those processes in the rubric. Moreover, one expert made a comment regarding the following sentence: When plants and animals die, their bodies decay bringing the carbon into the ground. She pointed out “actually, much the carbon goes back out as gas after decomposition, not “into the ground”. The content descriptions utilized in the rubric were elaborated based on the feedback from the content expert.

5) Rater training and scoring: two raters with biology and chemistry background received a total of four hours of training led by me. The training included an overview of the project, explanation of the scoring guide and rubrics, and discussion of sources of bias in

scoring with pre-scored student samples. During the training period, the two raters individually scored the items and compared their scores, and they then discussed any discrepancies in their scores until a consensus was reached for all scores. This whole process is called “norming” (Bresciani et al., 2004). Norming is the process of ensuring that raters understand the rubric in a consistent manner. During five weekly meetings, the rubrics were further revised through the norming processes.

6) Improving the rubric based on scoring: During the norming process, if an item had less than an 80% inter-rater agreement, the rubrics were refined to provide a clearer description for the raters and then the item was rescored with the revised version of the rubrics to achieve higher reliability. The final rubrics developed are presented in Appendix B.

7) Inter-rater reliability: Inter-rater reliability refers to the level of agreement to which two or more raters give the same rating to the same item. This study provides the two most common ways to quantify inter-rater reliability. One basic way is to calculate percentage of agreement between two raters. Although some would argue that percentage of agreement is not the best measure (Hayes & Hatch, 1999), some scholars prefer the percentage agreement measure over other correlation measures since it is simpler and easier to compute (Hayes & Hatch, 1999). To obtain the percentage agreement, first, I calculated the number of ratings that are in agreement and the total number of ratings, and finally converted the fraction to a percentage. As another option, inter-rater reliability for ordinal scale can be calculated as the intra-class correlation (ICC) (Shrout & Fleiss,

1979). ICC is an estimate of the variation among individuals (i.e., raters in this study) to the total measurement variance. The general formula for the ICC is thus

$$ICC = \frac{V_b}{V_T} = \frac{V_b}{V_b + V_e}$$

in which V_b = variance between raters, V_T = total variance, and V_e = error variance.

In order to choose the appropriate model of the ICC, there are two guidelines: (1) Do you have consistent raters to score all examinees' responses? (2) Do you have a sample or population of raters? This study allowed two raters to score all students' answers and assumed that the raters are "a sample". Based on these determinations, a "two-way mixed" model was used in this study. The "two-way mixed" model 1) considers both an effect of rater and of ratee (i.e., two effects) and 2) assumes that both are drawn randomly from larger populations (i.e., a random effects model). The range of possible values of the ICC is theoretically from 0 to 1 with 1 indicating perfect agreement, and 0 indicating poorer than chance agreement. According to Fleiss (1986), the cut off value of acceptable ICC is 0.75, which represents a good agreement.

3.6. Measurement model: two-parameter logistic model (2PLM) and generalized partial credit model (GPLM)

The final building block of the construct modeling is the measurement model. This model describes a way to relate the scored outcomes from the item design and the outcome space back to the construct map (Wilson, 2005). This study employs item response theory (IRT) as a measurement model to provide a framework in obtaining a valid and reliable assessment tool. Item response theory (IRT) is a powerful psychometric

technique that is used in education and psychological testing to develop and validate test data on both the item and test level (Lord, 1980). The idea of IRT is derived from the probability of each response as a function of the latent trait (i.e., examinees' ability) and item parameters that characterize the items. IRT consists of a set of mathematical models that use a latent trait (θ) and item parameters (e.g., difficulty, discrimination, and guessing). The ability value that has the highest likelihood becomes the ability estimate. IRT also generates statistical properties of item fits as well as statistical properties of the whole test fit. When the model is appropriate and the estimates of the item parameters are reasonably accurate, IRT suggests that the measure has appropriate construct validity (Hinkin, Tracey, & Enz, 1997).

IRT is based on a set of fairly strong assumptions, unlike classical test theory (CTT). If the assumptions are not met, the validity of the IRT estimates is severely compromised. IRT requires that the construct being measured in the assessment is unidimensional. Another important assumption is local independence (LD). LD assumes that an item response to one question is not contingent on a response to another question (Embretson & Reise, 2000), which provides us with statistically independent probabilities for item responses. When the IRT model satisfies the assumption of unidimensionality and LD, the latent trait estimates are not test-dependent, and item parameters are not sample-dependent (Yang & Kao, 2014). Another assumption under IRT is monotonicity, which is best displayed on a graph as a curve shaped like 'S' between the person ability level on the X-axis and the probability of a correct response to an item on the Y-axis. This graph is called an item characteristic curve (ICC). The ICCs reflect the monotonic

relationship between the level of probability of getting an item right and the level of the students' ability (Yang & Kao, 2014).

Selection of IRT models can be determined based on the number of scored responses and sample size. In this study, the responses are a combination of dichotomous items and polytomous items and the sample size is 454 individuals. Recommendations for the minimum sample size for reasonable estimation of model parameters range to 200 (Wright & Stone, 1979) for a one-parameter logistic model (1-PLM) to 350+ for a two-parameter logistic model (2-PLM) (Embretson & Reise, 2000) to 1000 for a three-parameter logistic model (3-PLM) (Lord, 1968). Thus, this study used a mixed-format IRT model: a two-parameter logistic model (2-PLM; Birnbaum, 1968) and a generalized partial credit model (GPCM; Muraki, 1992). The 2-PLM is used for dichotomous score items and the GPCM is used for the CR items with two or more score categories. The two models predict the probability of a correct response to an item based on ability and two item parameters, difficulty and discrimination. The item difficulty parameter (b), describes how difficult the item is, while the item discrimination parameter (a), determines how well an item identifies examinees with different levels of the latent trait (Embretson & Reise, 2000). Difficult items have large, positive theta values, whereas easy items have large, negative theta values (Reise & Waller, 2002). The theoretical range of the discrimination values is $-\infty$ to $+\infty$; however, items with negative discrimination values are considered problematic. The negative values indicate that examinees with high level of ability are less likely to answer items correctly.

The item response function of the 2-PLM is defined as

$$P_i(\theta) = \frac{\exp [a_i(\theta - b_i)]}{1 + \exp [a_i(\theta - b_i)]},$$

Where a_i is the discrimination parameter for item i , b_i is the difficulty, and θ is the ability of a person.

Under the GPCM, the probability that will respond in category x for item i with m_i+1 categories is expressed as:

$$P_{ix}(\theta) = \frac{\exp [\sum_{k=0}^x a_i(\theta - b_{ik})]}{\sum_{h=0}^{m_i} \exp [\sum_{k=0}^x a_i(\theta - b_{ik})]},$$

Where a_i is the discrimination parameter for item i and b_{ik} is the step difficulty parameter for category k .

3.7. Data analyses

As a preliminary check on the data, demographic differences in ISA scores were examined (see Table 8). The next step was to evaluate IRT's assumptions of unidimensionality and local independence. Both an exploratory factor analysis (EFA) and a confirmatory factor analysis (CFA) can be used to determine the dimensionality (i.e., number of factors) for the item responses in a scale. If a factor analysis identifies a single dimension (or factor), then the assumption of unidimensionality is met. An EFA is the most appropriate when hypotheses about a scale's internal structure are few. For example, if we began our discussion of the 19 items by assuming that we had no idea about dimensionality reflected in those items, an EFA is preferable to a CFA. In contrast, a CFA is useful when there are clear hypotheses about a test's dimensionality. In other words, a CFA is appropriate to examine a test's dimensionality when a test developer has

clear expectations about the number of factors or dimensions underlying the items, the links between the items and factors, and the association between the factors (Furr & Bacharachp, 2008).

There are many views in the literature on how unidimensionality can be decided in an EFA. For instance, Reckase (1979) believed that if the first order factor explains over 20% of the variance of a set of items, the measure is unidimensional. Lord (1980) argued that if the ratio of eigenvalue of the first factor to that of the second is large and the second eigenvalue is not larger than any of the others, then the measure has unidimensionality. In order to check unidimensionality in a CFA, the fit indices of a one-factor model should satisfy the cutoff criteria. The Mplus program (Muthén and Muthén, 1998-2015) provides a chi-square (X^2) test, Root Mean Square Error Approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and Weighted Root Mean Square Residual (WRMR). The chi-square (X^2) significance test should be non-significant. The CFI and TLI values should be greater than 0.9; while the RMSEA should be less than 0.05 to be an adequate one-factor model (Hu & Bentler, 1999; Kline, 2015). Yu (2002) recommends a cutoff value of WRMR <0.95 or 1.0.

To examine the assumption of local independence, the residual correlation matrix from a CFA was tested. If there are any residual correlations between two items that exceed the absolute value of 0.20, this indicates local dependence (Lorber et al., 2014). Local dependence in IRT models was also evaluated by standardized local dependence (LD X^2) statistics in which all values of LD X^2 greater than 10 indicated the presence of local dependence (Chen & Thissen, 1997). A violation of the local independence

assumption leads to overestimation of the reliability and test information function, and inappropriate standard error estimates of items (Sireci et al., 1991). All factor analyses were conducted, using Mplus software version 7.4 (Muthén & Muthén, 1998-2015).

Main IRT analyses were carried out to obtain detailed psychometric properties of item fit, item difficulty, item discrimination, differential item functioning (DIF) with regard to gender using IRTPRO 3 (Scientific Software International, 2016). Item fit statistics, $S-X^2$, express the degree of fit or misfit between observed and expected values in the data (Orlando & Thissen, 2000, 2003). The fit statistic was used to assess if each item fits the model. P-value, associated with the $S-X^2$ fit statistics, indicates an adequate fit of the item if it is greater than 0.05 (Sabbag & Zieffler, 2015). If most of the items fall within the acceptable range, the assessment model fits well. For construct validity of the assessment items, item parameter estimates (difficulty and discrimination) and standard error of the estimates obtained from both 2-PLM and GPCM were provided.

This study provides the item information curve (IIC) of each item and a test information curve (TIC). The IIC is a graphical summary that reflects how well the item measures a latent construct. For example, items with greater information are better at discriminating among students at different levels of the trait. Item information can vary at different levels of the trait rather than providing a single estimate of the item-trait relation. Moreover, IICs can be summed to generate a test information curve (TIC). A TIC shows how well the test is doing in estimating ability over the whole range of ability scores (Baker, 2001) and is very crucial for test development and item analysis.

Additionally, a facet map of the distribution of persons, items, and interdisciplinarity was generated, using the Facets 3.71.4 program (Linacre, 2015). The facet map is a graphical representation linking the item difficulty, interdisciplinarity of items, and person ability estimates on the common scale in logits. Specifically, unlike a Wright map, the facet map allows qualitative comparisons on the level of difficulty among interdisciplinary and disciplinary items. The larger the logit score for the items, the more difficult the items are.

Fairness of assessment is an important building block in the process of assessment validation (AERA et al., 2014). A fair test provides examinees with an equal opportunity to demonstrate their knowledge and ability relevant to the purpose of the assessment. That is, examinees with similar abilities on a test should logically show similar performance on individual items regardless of their gender, culture, ethnicity, or race (AERA et al., 2014). In order to investigate the assessment fairness, differential item functioning (DIF) can be used to check group differences in assessment performance. Specifically, DIF compares the probability to get an item right in one group to the probability to obtain a right answer in another group for examinees with similar abilities. This study used Lord's χ^2 Wald test (1980) for detecting DIF of gender and race variables.

CFA is a commonly used to verify the underlying structure of the instrument (i.e., structural validity) and to quantify the relationship between each item and the construct (Salkind, 2010). Thus, the use of CFA allows for the comparison of proposed factor structures of an instrument and for determination which model provides the best

representation of the underlying structure of the ISA. This study assumed that the ISA items could be divided into interdisciplinary and disciplinary in the item development stage, and confirmed which items will load onto which factors through content experts' interviews. 13 items were specified to load on an interdisciplinary factor and other 6 items were loaded on a disciplinary factor (see Figure 6). This study proposed three hypothesized models for confirmation of the possible latent factors: a two-factor model with orthogonal factors, a two-factor model with oblique factors, and a second order model. Model fit was assessed with several fit indices, including a chi-square (X^2) test, Root Mean Square Error Approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and Weighed Root Mean Square Residual (WRMR). To further examine the structural validity of the ISA, I also examined the pattern of cross-scale correlations by Point-Biserial correlation coefficients for all items and the ISA subscales.

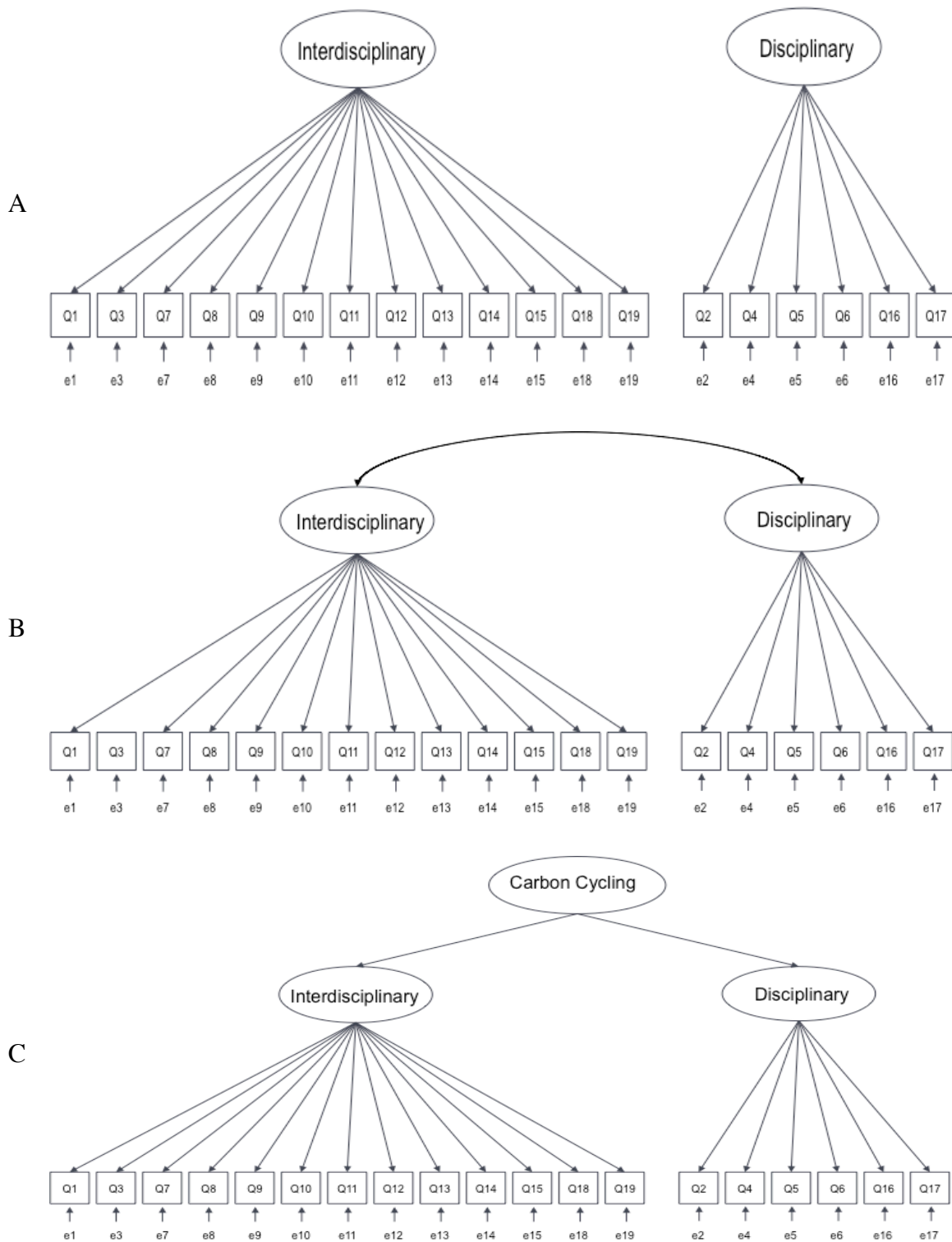


Figure 6. Three hypothesized structural models of the ISA, A: Two- factor model with orthogonal factors, B: Two-factor with oblique factors, C: Second order model.

This study includes internal consistency (i.e., Cronbach's alpha) and inferential statistical analyses (t-test or ANOVA). I used student ability estimates (theta) from IRT instead of traditional raw scores as a dependent variable. The relationship between four groups with varying numbers of science courses taken and the students' ISA scores was identified using ANOVA. In a similar manner, the difference in the ISA scores according to grade levels was examined with ANOVA. Also, the ISA score difference of females and males was identified using a t-test. The relationship between interdisciplinary score and disciplinary score was identified with Pearson correlation. The assumptions for ANOVA and t-test, normality and homogeneity of variances, were checked before performing the all analyses.

Chapter 4: Results

This research focuses on the development of the ISA (Interdisciplinary Science Assessment) and its validation. This chapter presents the results of the entire process of item development and the evidence of the construct validity of the ISA. Section 4.1 describes the qualitative process of item development. Section 4.2 shows the process of item modification based on evidence from the content validation and the pilot test. For outcome space stage, section 4.3 discusses scoring rubric development and scoring strategies for the eight CR items. Lastly, section 4.4 highlights psychometric specifications to suggest evidence for the construct validity of the ISA using IRT models, and provides the result of internal consistency and inferential statistics with students' demographic information.

4.1. Development of the construct map

A core principle of construct modeling is starting with the development of a construct map, not item creation. This study developed a construct map of interdisciplinary understanding of carbon cycling. The map consists of five levels of interdisciplinary understanding in a hierarchical fashion. Table 5 shows the construct map with specific responses regarding deforestation. This construct map was developed by referring to the KI (Knowledge Integration) and interdisciplinary learning literature (e.g., Golding, 2009; Liu et al., 2008).

Table 5. Construct map of the interdisciplinary understanding of carbon cycling (adapted from Golding, 2009 & Liu et al., 2008).

| Interdisciplinary understanding level | Description | Response example on deforestation |
|--|--|--|
| <i>Fully Interdisciplinary understanding</i> | A student uses relevant scientific concepts and principles to explain a specific event in carbon cycling to demonstrate reasoned interdisciplinary understanding. | Deforestation firstly produces a lot of CO ₂ since many trees are burned, and CO ₂ is released by machines that help cut down trees. Additionally, since there are less trees, less CO ₂ is being converted back to oxygen. With less trees, the CO ₂ in the atmosphere will increase. And by burning the trees, more CO ₂ will enter the atmosphere. |
| <i>Partially interdisciplinary understanding</i> | A student response has a partially accurate understanding of the scientific concepts and principles based on more than two necessary disciplines to explain a specific event in carbon cycling. | CO ₂ is used by trees for photosynthesis. That CO ₂ is stored within the entire tree. When the tree is cut down, the CO ₂ is released from the decomposing tree. That CO ₂ will seep in the soil, or escape into the atmosphere. |
| <i>Unidisciplinary understanding</i> | A student response has a partially accurate understanding of the scientific concepts and principles and his/her answer is grounded in a single necessary discipline to explain a specific event in carbon cycling. | The more deforestation the less there are plants to perform photosynthesis to break down the CO ₂ which means it will increase in the atmosphere. |
| <i>No response or Irrelevant</i> | A student response is blank, not relevant to scientific phenomenon represented in the item, or shows only the restatement of the prompt. | Deforestation could lead to an imbalance because this would lead to a decrease in CO ₂ levels. When there is a reduction in number of trees there will be a reduction in CO ₂ levels. Trees create CO ₂ and with less trees the production of it will be less. |

4.2. Item design

Selection of core concepts in carbon cycling. Carbon cycling contains a broad range of scientific concepts (e.g., global warming, deforestation); thus, it is essential to narrow and select the core concepts to determine the range of content tested. This study used two ways to do this. One is to refer to the performance expectations of the disciplinary core ideas (DCIs) in the NGSS. The DCIs are organizing key concepts of a single discipline

for understanding more complex issues and ways to solve them. The DCIs, however, are interdisciplinary with connections across multiple sciences. I created Figure 7, which reflects the connections that are closely intertwined between Physical Sciences, Earth and Space Sciences, and Life Science.

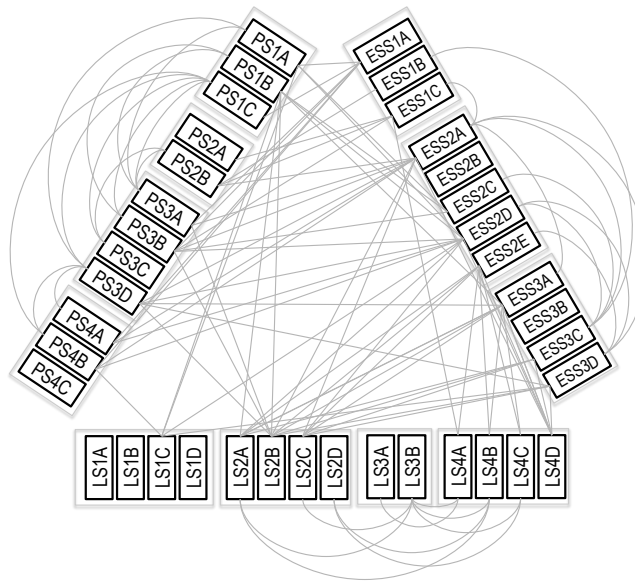


Figure 7. Interdisciplinary connections in the NGSS (PS: Physical Sciences, ESS: Earth and Space Sciences, LS: Life Sciences).

By matching what is assessed to the expectations provided by the NGSS, students will know what they should be learning and what will be tested. This study explored the performance expectations for middle and high school students related to carbon cycling in the NGSS. For example, photosynthesis and cellular respiration are important concepts in the life science area at both the middle school and high school levels. The excerpts from the NGSS show examples of statements of what students should know and what they should be able to do in order to learn photosynthesis and cellular respiration.

- The chemical reaction by which plants produce complex food molecules (sugars) requires an energy input (i.e., from sunlight) to occur. In this reaction, carbon dioxide and water combine to form carbon-based organic molecules and release oxygen. (Secondary to MS-LS1-6)
- Cellular respiration in plants and animals involve chemical reactions with oxygen that release stored energy. In these processes, complex molecules containing carbon react with oxygen to produce carbon dioxide and other materials. (Secondary to MS-LS1-7)
- Photosynthesis and cellular respiration are important components of the carbon cycle, in which carbon is exchanged among the biosphere, atmosphere, oceans, and geosphere through chemical, physical, geological, and biological processes. (HS-LS2-5)

Additionally, the ESS3 Disciplinary Core Idea from the NGSS framework includes an important point, “human activities” in carbon cycling. Knowledge of human activities helps students formulate an answer to questions such as: “How do human activities affect other Earth systems and living organisms in those, and how does the release of excess CO₂ from burning fossil fuels have an impact on global climate?” In the ESS3 performance expectations, students are expected to demonstrate their understanding of one of the core ideas in carbon cycling by constructing an explanation and engaging in an argument.

- Human activities have significantly altered the biosphere, sometimes damaging or destroying natural habitats and causing the extinction of other species. But

changes to Earth's environments can have different impacts (negative and positive) for different living things. (MS-ESS3-3)

- Human activities, such as the release of greenhouse gases from burning fossil fuels, are major factors in the current rise in Earth's mean surface temperature (global warming). Reducing the level of climate change and reducing human vulnerability to whatever climate changes do occur depend on the understanding of climate science, engineering capabilities, and other kinds of knowledge, such as understanding of human behavior and on applying that knowledge wisely in decisions and activities. (MS-ESS3-5)
- Changes in environmental conditions (e.g., deforestation, global warming) may result in: (1) increases in the number of individuals of some species, (2) the emergence of new species over time, and (3) the extinction of other species. (HS-LS4-5)

Another important sub-core idea in carbon cycling is “matter cycles and energy flow” in living organisms. The performance expectations in LS2, PS3, and ESS2 from the NGSS show that students need to develop their qualitative ideas on the process of matter cycles and energy transfer. Students develop their understanding of important qualitative ideas about the concept of transfer of energy from one object or system of objects to another. Also, students are expected to understand the movement of matter among plants, animals, and decomposers, through the food chain or food web where carbon is a very important element in matter cycles.

- Food webs are models that demonstrate how matter and energy are transferred

between producers, consumers, and decomposers as the three groups interact within an ecosystem. Transfers of matter into and out of the physical environment occur at every level. Decomposers recycle nutrients from dead plant or animal matter back to the soil in terrestrial environments or to the water in aquatic environments. (MS-LS2-3)

- Plants or algae form the lowest level of the food web. At each link upward in a food web, only a small fraction of the matter consumed at the lower level is transferred upward, to produce growth and release energy in cellular respiration at the higher level. Given this inefficiency, there are generally fewer organisms at higher levels of a food web. Some matter reacts to release energy for life functions, some matter is stored in newly made structures, and much is discarded. The chemical elements that make up the molecules of organisms pass through food webs and into and out of the atmosphere and soil, and they are combined and recombined in different ways. At each link in an ecosystem, matter and energy are conserved. (HS-LS2-4)
- Energy cannot be created or destroyed, but it can be transported from one place to another and transferred between systems. (HS-PS3-1), (HS-PS3-4)
- All Earth processes are the result of energy flowing and matter cycling within and among the planet's systems. This energy is derived from the sun and Earth's hot interior. The energy that flows and matter that cycles produce chemical and physical changes in Earth's materials and living organisms. (MS-ESS2-1)

Beyond searching the NGSS content framework, this study identified the carbon cycling core concepts from content experts' concept maps. Although there are many ways to elicit and reveal their ideas, a concept map provides a good means for generating ideas and their interconnections graphically. This study went through the experts' concept maps and analyzed their perceptions to identify subtopics of carbon cycling. Frequently used terms or phrases were assumed to be key concepts. The examples of concept maps are shown in Figure 8.

Experts' common terms for carbon cycling were "photosynthesis," "cellular respiration," and "human activities" such as burning fossil fuels.

A professor in the biology department emphasized the concepts of photosynthesis by plants or bacteria, respiration of animals, plants, and microbes, and the changes of atmospheric CO₂ levels. He demonstrated his knowledge of various forms of carbon among different reservoirs in the Earth system. For example, he mentioned that carbon is incorporated into sediments and it is in water in the form of carbonic acid. He also included human activities of burning fossil fuels and farming which represent the effects of increased atmospheric CO₂ and CH₄ levels.

Another professor in the biology department mentioned that carbon atoms can be in either an abiotic system or a biotic one. In an abiotic system, carbon atoms can be a part of inorganic minerals in rocks and CO₂ molecules in the atmosphere. In a biotic system, he incorporated the concept of photosynthesis and linked it to the concept of cellular respiration including light reaction, Calvin cycle, and glycolysis. He also mentioned the role of decomposers, releasing CO₂ and using the products (e.g., sugars) of

photosynthesis. He also stated that organisms consist of sugars or cellulose, and they can eventually form into fossil fuels, which in turn can be burned, releasing CO₂ into the atmosphere.

A Ph.D. student in the science education program also demonstrated her knowledge of photosynthesis and respiration, and decomposition processes like other experts. She drew the processes in detail. She showed that CO₂ can be released into the atmosphere through three processes: decomposition, cellular respiration, and burning of fossil fuels. She mentioned various carbon reservoirs such as water, soil, air, and organisms, specifically describing the formation of carbonic acid after CO₂ dissolves in water and the acidification of water. She also showed that CO₂ and H₂O, the product of cellular respiration, are greenhouse gases that hold heat in the air and cause global warming.

A professor in the geoscience department described the four major reservoirs of carbon and how they are interconnected with each other and how the carbon exchange between the reservoirs occurs as the result of various biological, chemical, geological, and physical processes. He showed the photosynthesis and respiration processes between atmosphere and biosphere, or atmosphere and hydrosphere, and phenomena such as volcanic eruptions and fossil fuel burning between geosphere and atmosphere.

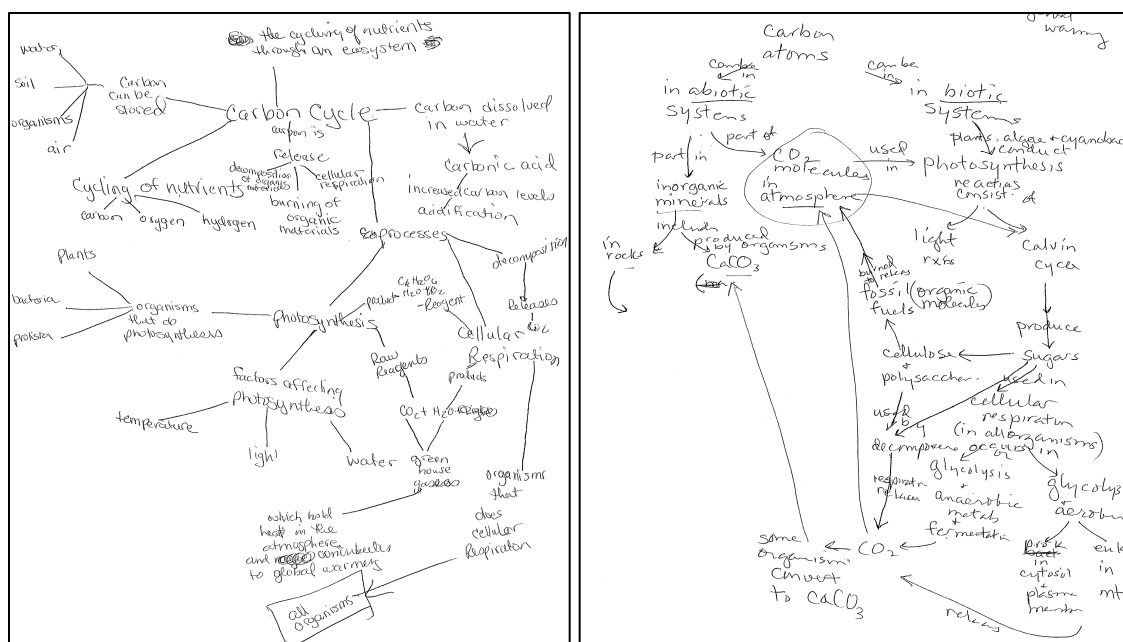


Figure 8. Concept map examples of the content experts.

Based on the experts' concept maps and an NGSS content framework on performance expectations, nine core themes for the content selection in carbon cycling were selected. They were photosynthesis, cellular respiration, decomposition, carbon reservoirs, fossil fuels, deforestation, food chain (movement of carbon and energy flow), ocean acidification, and global warming. Based on the nine core concepts identified as relevant, an initial item pool was developed.

Item format. The ISA used both multiple choice (MC) items and construct response (CR) items. MC items are the preferred and common assessment format for assessing students' knowledge and skills. MC items have key advantages like ease of administration, objective and high-speed test scoring, and the low cost of scoring (Roediger & Marsh, 2005). Although CR items basically require a longer administration time and more scoring effort, they provide students with a chance to express or elaborate

their ideas and knowledge, especially when high order thinking, e.g., interdisciplinary understanding and application of knowledge, is assessed in a test. Taking into consideration all of these elements, the preliminary item pool consisting of 20 items (11 MC items and 9 CR items) was developed.

Content validity. Content validity is the determination of the content representativeness or content relevance of the items. Content analysis occurred in three phases in this study: (1) initial assessment development through examination of the NGSS standards and experts' concept maps, (2) review after the assessment had been constructed, and (3) evaluation of unidimensionality in IRT models or factor analyses after the final version of the assessment had been administered to students.

During Phase 1, evidence of the content being assessed should be provided to see whether the content is consistent with the intent of the content standards. Also, the assessment should clarify the intended construct and the relationship between the construct and the items (Miller & Linn, 2000). Phase 1 formulates answers to questions such as: "Does the ISA fully represent the interdisciplinary construct, as defined by the test? Are the items aligned with relevant content in the NGSS or experts' content framework?" The individual interviews with the nine content experts provided information about the interdisciplinarity of the items. For example, all experts agreed that the item regarding ocean acidification below has interdisciplinarity. The item contains concepts from multiple science disciplines: movement of carbon between atmosphere and hydrosphere as a part of the carbon cycle (Earth science), CO₂ production from fossil fuel combustion and Henry's law, the impact of calcification rates in lowering pH in the

ocean (chemistry), and CaCO_3 needed for certain coral reefs to make their structures (biology).

Item 8) Scientists have recently calculated that approximately 26% of all CO_2 emitted from human-related activities, such as the combustion of fossil fuels, was absorbed by oceans during the decade 2002-2012. This resulted in 2.5 billion gigatons of excess carbon moved from the atmosphere into the ocean each year over the course of a decade. Scientists are concerned that the mass of CaCO_3 deposited annually in coral reefs is decreasing. They expect that in 2050 the total amount of CaCO_3 in coral reefs will be 20 percent less than it is currently.

Use this information to describe how the combustion of fossil fuels might affect the loss of coral reefs in ocean water, even though the amount of CO_2 in the atmosphere is increasing.

This study developed both interdisciplinary and unidisciplinary items. The other item about ocean acidification requires only unidisciplinary understanding from chemistry. The item example is as follows:

Item 17) Ocean acidification is caused when the ocean absorbs more CO_2 , resulting in:

- A) A decrease in the pH of ocean water
- B) An increase in the pH of ocean water
- C) A decrease in temperature of ocean water
- D) An increase in temperature of ocean water
- E) An increase in methane emissions

Phase 2 shows judgments of the content experts regarding the representativeness and adequacy of an individual prompt for subdomains of the content and construct. Content experts' responses were used to calculate the content validity index (CVI). Table 6 presents the relevance ratings of nine experts for 20 items. Only 18 out of 20 items received relevance ratings of 3 or 4 by all experts. According to the minimum criteria for the CVI proposed by Lynn (1986) (see Table 3 in chapter 3), item CVI should be higher than 0.78. In other words, seven out of nine experts needed to choose relevance ratings of

3 or 4. As shown below in Table 6, out of the 20 items, the CVI of 18 items was greater than 0.78 and two items failed to meet this criterion.

Table 6. CVI ratings by experts on 20 items.

| Item | Expert 1 | Expert 2 | Expert 3 | Expert 4 | Expert 5 | Expert 6 | Expert 7 | Expert 8 | Expert 9 | Number in Agreement | Item CVI |
|------|----------|----------|----------|----------|----------|----------|----------|----------|----------|---------------------|----------|
| 1 | X | X | X | X | X | X | X | X | X | 9 | 1.00 |
| 2 | X | X | X | X | X | X | X | -- | X | 8 | 0.83 |
| 3 | X | -- | X | X | X | X | X | X | X | 8 | 0.83 |
| 4 | X | X | -- | X | X | -- | -- | X | X | 6 | 0.67 |
| 5 | X | -- | X | X | X | X | X | -- | X | 7 | 0.78 |
| 6 | X | X | X | X | X | X | X | X | X | 9 | 1.00 |
| 7 | X | -- | X | X | X | X | X | X | X | 8 | 0.83 |
| 8 | X | X | X | X | X | X | X | X | X | 9 | 1.00 |
| 9 | X | X | X | X | X | X | X | X | X | 9 | 1.00 |
| 10 | X | X | X | X | X | X | X | X | X | 9 | 1.00 |
| 11 | X | X | X | X | X | X | X | X | X | 9 | 1.00 |
| 12 | X | X | X | X | X | -- | -- | X | X | 7 | 0.78 |
| 13 | X | -- | X | X | X | X | -- | X | X | 7 | 0.78 |
| 14 | X | X | X | X | X | X | X | -- | X | 8 | 0.83 |
| 15 | X | X | X | X | X | X | X | X | X | 9 | 1.00 |
| 16 | X | X | X | X | X | -- | X | X | X | 8 | 0.83 |
| 17 | X | X | X | X | X | X | X | X | X | 9 | 1.00 |
| 18 | X | X | X | X | X | X | X | X | X | 9 | 1.00 |
| 19 | X | X | X | X | X | X | X | X | X | 9 | 1.00 |
| 20 | X | X | -- | X | X | X | -- | X | -- | 6 | 0.67 |

Note: X: items rated 3 or 4 on a 4-point Likert type scale

Since two items (item 4 and 20) did not achieve the required agreement from the experts, they were eliminated. Based on the feedback received from the experts, other items were modified by (1) changing from a CR item to a MC item, (2) adding phrases to clarify the meaning of the questions, (3) removing extraneous information in prompts, or (4) adding two new items pertaining to the physics and earth science disciplines. This led to the second version of the instrument and it was used in the pilot test.

Phase 3 will be shown in section 4.4, which discusses the results of data analyses.

The fit indices in IRT models are used to check the relevance of the intended test construct. Misfitting items show a possibly different and irrelevant construct. A facet map

also allows for verification of the representativeness of test content through identifying the level of item difficulties and the ability of examinees within a range (Smith, 2001).

The second version of the ISA (20 items), following the confirmation process of content validity, was pilot tested. After the pilot test, only one item was eliminated because the item required simple recall of ideas but was shown to be the most difficult for students. The item is about the largest carbon reservoir in the Earth. The item deleted is as follows.

Q) Most of the Earth's carbon resides in

- A. I don't know.
- B. Soils and vegetation
- C. The ocean
- D. Sedimentary rocks
- E. The atmosphere

4.3. Inter-rater reliability

Based upon a review of the literature on rubric design (Wiggins, 1998), this study adopted a multi-step approach. Two coders graded 454 students' responses. The details of the steps are illustrated in chapter 3. To establish the inter-rater reliability of the coded data set, the percentage agreement and the intra-class correlation coefficients (ICCs) based on a two-way mixed model were evaluated.

Table 7. Percentage agreement and intra-class correlation coefficients assessing inter-reliability.

| Item | Percentage agreement (%) | ICC |
|------|--------------------------|---------|
| 2 | 92.51 | 0.991** |
| 4 | 94.71 | 0.989** |
| 7 | 91.63 | 0.987** |
| 8 | 94.05 | 0.996** |
| 12 | 92.51 | 0.992** |
| 14 | 90.53 | 0.993** |
| 15 | 90.09 | 0.980** |
| 19 | 92.95 | 0.994** |

Note: ** Significant at 0.001 level

Table 7 summarizes the percentage agreement and ICCs. The percentage agreement between raters for all eight CR items was greater than 90%, ranging from 90.09%-94.7%. Disagreements were resolved by group consensus in weekly meetings. The ICCs for individual items ranged from 0.980 (item 15) to 0.996 (item 8). The average ICC for the 8 CR items was 0.990 with all items demonstrating excellent inter-rater reliability; thus, the items were identified as having statistically significant levels of inter-rater reliability. Based on the results, the rubrics are considered reliable to evaluate students' performances.

4.4. Results of data analysis

This section outlines data analyses used to examine the final version of the ISA. First, the descriptive statistics were reported. Second, IRT assumptions (unidimensionality and local independence) were examined. Third, the analyses of two IRT models were conducted. Fourth, internal consistency of the ISA was assessed. Finally, performance information through statistical analyses of selected demographic characteristics of the students was provided.

Descriptive statistics. Descriptive statistics according to the demographic information

are presented in Table 8 below. These descriptive statistics, particularly the sample size and percent, mean, and standard deviation (SD), were measured. The total possible score for the instrument was 71 points. There are 11 MC items and each MC item was worth 1 point. The mean of the MC items was 5.93 out of 11. The assessment contains a total of 8 CR items. Among those, two CR and disciplinary items were worth 6 points and six interdisciplinary items were worth 8 points each. As disciplinary items require only one discipline's knowledge to answer correctly, the assigned score was lower than for interdisciplinary items. The mean of the CR items was 22.52 out of 60. The performance on CR items was lower than on the MC items. Also, the test can be categorized as 6 disciplinary items and 13 interdisciplinary items. The mean of the disciplinary items was 7.90 out of 16 while the mean of the interdisciplinary items is 20.56 out of 55.44. High school students in particular showed very low achievement with a mean of 11.39 compared to college/graduate students who had a mean of 30.29. Female and male students showed similar performance in the ISA, indicating only a 0.03-point difference. According to race demographics, Asian students had the highest average score with 31.58, followed by White and Hispanic.

Table 8. Demographic information and descriptive statistics (N=454).

| | N | Percent | Min | Max | Mean | SD |
|-------------------------------------|-----|---------|-----|-----|-------|-------|
| Item (0-71) | | | | | | |
| MC (0-11) | 11 | 57.9 | 0 | 11 | 5.93 | 2.32 |
| CR (0-60) | 8 | 42.1 | 0 | 45 | 22.52 | 10.66 |
| Disciplinary (0-16) | 6 | 31.6 | 0 | 16 | 7.90 | 3.30 |
| Interdisciplinary (0-55) | 13 | 68.4 | 0 | 43 | 20.56 | 9.54 |
| Overall | 19 | 100 | 4 | 56 | 28.46 | 12.09 |
| Grade level | | | | | | |
| High school | 44 | 9.6 | 4 | 46 | 11.39 | 9.01 |
| College/Graduate | 410 | 90.4 | 4 | 56 | 30.29 | 10.89 |
| Overall | 454 | 100 | 4 | 56 | 28.46 | 12.09 |
| Gender | | | | | | |
| Female | 264 | 58.1 | 4 | 55 | 28.47 | 11.50 |
| Male | 190 | 41.9 | 4 | 56 | 28.44 | 12.89 |
| Overall | 454 | 100 | 4 | 56 | 28.46 | 12.09 |
| Race | | | | | | |
| White | 177 | 39.0 | 4 | 54 | 29.73 | 11.56 |
| American Indian or Alaska Native | 0 | 0 | | | NA | NA |
| Asian | 131 | 28.9 | 4 | 55 | 31.58 | 11.51 |
| Native Hawaiian or Pacific Islander | 2 | 0.4 | 17 | 24 | 20.50 | 4.95 |
| Black or African American | 23 | 5.1 | 4 | 40 | 24.70 | 10.91 |
| Hispanic | 106 | 23.3 | 4 | 56 | 23.81 | 12.99 |
| Other | 15 | 3.3 | 9 | 51 | 25.80 | 11.21 |

Unidimensionality. Understanding the internal structure of a test is one piece of validity evidence that can be used to support the intended inferences and uses of test scores. In addition, one of the assumptions of IRT models is unidimensionality. Prior to the main analyses, the unidimensionality assumption was evaluated for the scale, using both confirmatory factor analysis (CFA) and exploratory factor analysis (EFA).

In CFA, several well established indices and criteria were used to assess the goodness of fit of a single-factor model. These criteria included a norm chi-square statistic (χ^2/df). The chi-square statistics evaluate the extent that a proposed model varies from the data. Nonsignificant p values are ideal, but this is unrealistic because the chi-square value is driven too heavily by the sample size (Healey, 2012). Thus, for

controlling the effect of the sample size, it is recommended that the chi-square value should be divided by the degrees of freedom (χ^2/df), generating a normed chi-square value. Acceptable range of this value is from as high as 5.0 to as low as 2.0 (Hooper, Coughlan, & Mullen, 2008). The norm chi-square value was 1.61, indicating an excellent fit. Other fit indices, CFI= 0.932, TLI=0.924, RMSEA= 0.037 with a 90% confidence interval between 0.028 and 0.045, and WRMR=0.925 (see Table 9), indicate that the one-factor model fits the data well (Hu & Bentler, 1999; Kline, 2015). This result provides empirical evidence that a single latent trait sufficiently explains the item responses. To confirm the one factor model, a chi-square difference test was performed with a two-factor model. A chi-square difference test enables us to decide which model is more appropriate between one- and two-factor models. The χ^2 difference is 0.60, $\Delta df=1$ (see Table 10), which indicates the difference test is not significant. In this case, the one-factor model should be chosen according to a parsimonious rule (Kline, 2015).

Table 9. CFA model fit indices of ISA-one factor model.

| Model fit for the proposed model | χ^2 | <i>df</i> | <i>p-value</i> | Normed χ^2 (χ^2/df) | CFI | TLI | RMSEA (90% CI) | WRMR |
|----------------------------------|----------|-----------|----------------|---------------------------------|-------|-------|----------------------|-------|
| | 244.194 | 152 | < .001 | 1.61 | 0.932 | 0.924 | 0.037 (0.028: 0.045) | 0.925 |

Note: 90% CI=90% confidence interval of RMSEA.

Table 10. CFA model fit indices of ISA-two factor model.

| Model fit for the proposed model | χ^2 | <i>df</i> | <i>p-value</i> | Normed χ^2 (χ^2/df) | CFI | TLI | RMSEA (90% CI) | WRMR |
|----------------------------------|----------|-----------|----------------|---------------------------------|-------|-------|----------------------|-------|
| | 243.595 | 151 | < .001 | 1.61 | 0.932 | 0.923 | 0.037 (0.028: 0.045) | 0.925 |

Note: 90% CI=90% confidence interval of RMSEA.

Figure 9 depicts the one-factor model. The standardized factor loading of each item onto the latent trait (about carbon cycling) for the one-factor model was significant ($p < .001$) and ranged from 0.29 to 0.65. Factor loadings greater than 0.30 are considered an acceptable magnitude (Crocker & Aligina, 1986).

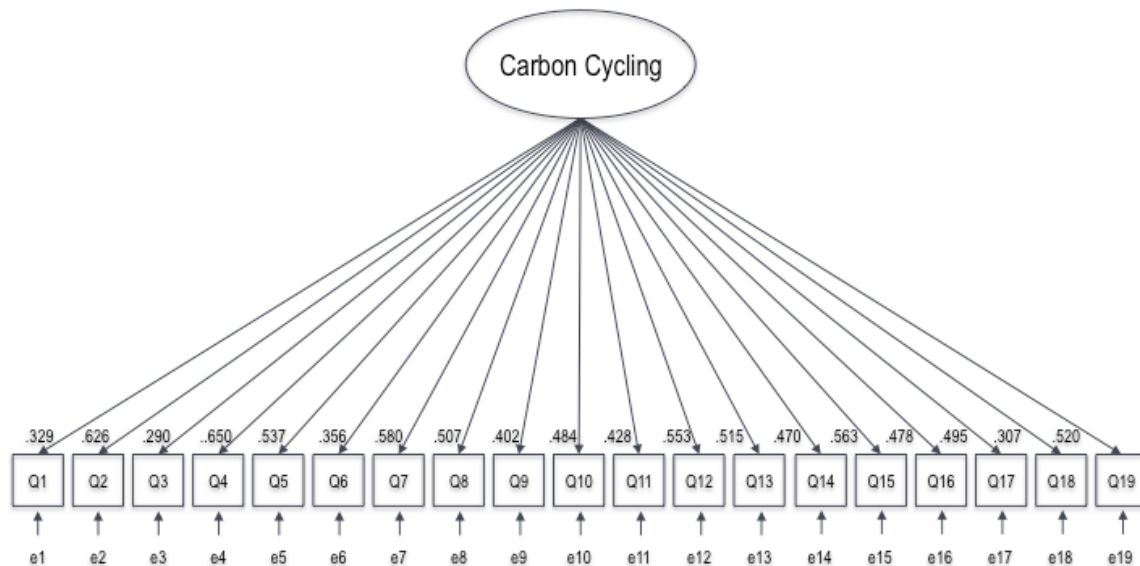


Figure 9. One factor CFA model of the ISA.

EFA can provide additional information of unidimensionality. It is suggested that the first factor accounting for at least 20% of the variance (Reckase, 1979) and having an eigenvalue that is four or five times greater than that of the second factor (Lord, 1980) is considered evidence for unidimensionality. The first factor extracted had an eigenvalue of 4.45, which was more than 3.32 times as large as the eigenvalue of 1.34 for the second factor. Also, the first factor accounted for 23.40% of the total variation. Even though the ratio of the first to the second eigenvalue is a little smaller than the criteria Lord recommended, a scree plot showed additional evidence of the dominance of the first factor (Figure 10), indicating sufficient unidimensionality.

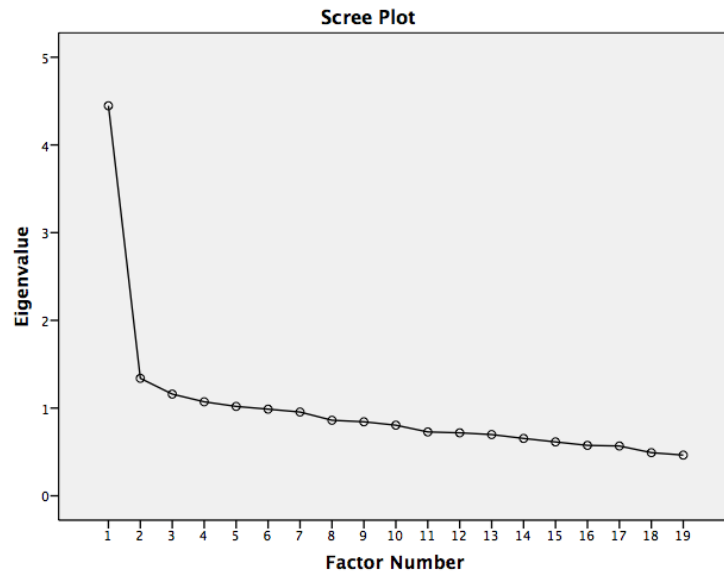


Figure 10. Scree plot of the ISA.

Local independence. Local independence (LD), the second important assumption of the IRT model, was assessed by the (approximately) standardized LD X^2 statistics (Chen & Thissen, 1997). The LD X^2 statistics can be obtained through computation by comparing the observed and expected frequencies in each of the two-way cross tabulations between responses to each item. These diagnostic statistics are standardized X^2 values, meaning that they are approximately z-scores. Z-scores are computed by subtracting the degrees of freedom from those X^2 -distributed statistics and dividing it by the square root of twice the degrees of freedom. Values of the LD X^2 were below the threshold of 10, which is considered positive evidence of local independence (Teresi et al., 2015). All LD X^2 statistics values of the ISA were relatively small, ranging from -1.6-5.3, which indicates no evidence of local dependence. The pair of items with the largest LD statistics (5.3) is item 13 and item 16.

Item fit. $S-X^2$ item-fit statistic suggested by Orlando & Thissen (2000, 2003) expresses the degree of fit or misfit in items. A statistically significant difference between observed and modeled values of $S-X^2$ statistics indicates misfit for items having p -values less than 0.05. The $S-X^2$ test for the item fit is shown in Table 11. The majority of the items fit well for the specified 2PLM and GPCM. Only item 5 among 19 items had a p -value less than 0.05, which indicates a misfit item.

Table 11. $S-X^2$ Item Level Diagnostic Statistics.

| Item | $S-X^2$ | df | p-value |
|------|---------|------|---------|
| 1 | 42.49 | 43 | 0.4945 |
| 2 | 238.40 | 143 | 0.5934 |
| 3 | 45.73 | 43 | 0.3586 |
| 4 | 121.99 | 126 | 0.5849 |
| 5 | 37.64 | 20 | 0.0098* |
| 6 | 51.05 | 44 | 0.2159 |
| 7 | 200.40 | 177 | 0.1097 |
| 8 | 174.87 | 155 | 0.1310 |
| 9 | 29.70 | 40 | 0.8837 |
| 10 | 36.37 | 39 | 0.5915 |
| 11 | 38.25 | 40 | 0.5504 |
| 12 | 149.56 | 158 | 0.6725 |
| 13 | 31.59 | 39 | 0.7949 |
| 14 | 177.16 | 167 | 0.2803 |
| 15 | 121.94 | 111 | 0.2246 |
| 16 | 38.40 | 42 | 0.6306 |
| 17 | 32.96 | 41 | 0.8105 |
| 18 | 43.66 | 43 | 0.4445 |
| 19 | 123.49 | 140 | 0.8387 |

* p -values < 0.05.

Item parameters. Under item response theory, the difficulty of an item describes where the item functions along the ability scale. For example, an easy item functions among the low-ability examinees and a hard item functions among the high-ability examinees; thus, difficulty is a location index. The item difficulties ranged from -2.74~0.99 across the MC

items. Item 18 is the most difficult item while item 5 is the easiest item (see Table 12 & Table 13). Items 1, 3, 5, 6, 13, 16, and 17 with negative item difficulty are easy items because the probability of a correct response is high for low-ability students. Items 2, 4, 7, 9, 10, 11, 12, and 14 have relatively medium difficulty because the probability of correct response is low at the lowest ability levels, around .5 in the middle of the ability scale and near 1 at the highest ability levels. Items 8, 15, 18, and 19 represented hard items. The probability of a correct response was low for most of the ability scale and increased only when the higher ability levels are reached. Even at the highest ability level shown (+3), the probability of a correct response was only 0.8 for the most difficult item.

Discrimination parameters describe how well an item can differentiate between examinees having abilities below the item location and those having abilities above the item location. The discrimination ranged from 0.30 (Item 14) to 1.08 (Item 5). Item 14 with low discrimination shows the probability of a correct response at low ability levels is nearly the same as it is at high ability levels. Item 5 had a high level of discrimination where the probability of a correct response changes very rapidly as ability increases.

Table 12. 2PLM item parameters estimates, logit: $a\theta + c$ or $a(\theta - b)$.

| Item | <i>Discrimination (a)</i> | <i>Difficulty (b)</i> |
|------|---------------------------|-----------------------|
| 1 | 0.55 | -0.73 |
| 3 | 0.46 | -0.76 |
| 5 | 1.08 | -2.74 |
| 6 | 0.59 | -1.00 |
| 9 | 0.65 | 0.31 |
| 10 | 0.93 | 0.99 |
| 11 | 0.74 | 0.86 |
| 13 | 0.91 | -0.40 |
| 16 | 0.80 | -0.46 |
| 17 | 0.90 | -0.40 |
| 18 | 0.50 | 1.53 |

Table 13. GPC model item parameter estimates, logit: $a[k(\theta - b) + \sum dk]$.

| Item | Discrimination (<i>a</i>) | Location parameter (<i>b</i>) | d_1 | d_2 | d_3 | d_4 | d_5 | d_6 | d_7 | d_8 | d_9 |
|------|--------------------------------|------------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 2 | 0.63 | 0.67 | 0 | 0.81 | 0.44 | 0.15 | 0.14 | 0.81 | 0.72 | -- | -- |
| 4 | 0.86 | 0.14 | 0 | 1.85 | 1.79 | 1.15 | 0.34 | 1.86 | 2.58 | -- | -- |
| 7 | 0.36 | 0.26 | 0 | 5.26 | 5.58 | 0.20 | 1.90 | 0.76 | 1.66 | 2.69 | 2.28 |
| 8 | 0.41 | 1.74 | 0 | 0.16 | 1.89 | 0.90 | 1.23 | 0.04 | 0.19 | 3.41 | 0.24 |
| 12 | 0.34 | 0.85 | 0 | 4.35 | 3.91 | 1.63 | 6.71 | 2.03 | 2.27 | 0.43 | 4.83 |
| 14 | 0.30 | 0.77 | 0 | 1.59 | 2.54 | 0.37 | 1.48 | 2.36 | 7.95 | 1.14 | 3.92 |
| 15 | 0.53 | 1.13 | 0 | 2.92 | 7.23 | 0.64 | 1.39 | 8.77 | 2.25 | 1.24 | 1.06 |
| 19 | 0.35 | 1.47 | 0 | 3.35 | 2.52 | 0.58 | 2.00 | 2.03 | 1.16 | 0.24 | 0.52 |

Note: *a*: slope parameter, *b*: item location characterizing overall difficulty, *d*: threshold parameters (relative difficulty of each step needed to transition from one category to the next in an item).

Item characteristic curve (ICC) and item information curve (IIC). Each item in the ISA has its own item characteristic curve. The *x*-axis of the ICC indicates increasing levels of ability required for an item from left to right on logit scale ranging from -3.00 to $+3.00$. The *y*-axis of the ICC indicates the probability of a response to the item. For example, Figure 11 presents the ICC of item 5. One misfit item, item 5, was very highly discriminating ($a=1.08$) as gauged by the steepness of the graph at its midpoint. This item sharply separates students with trait levels a little above and a little below the item difficulty of $b = -2.57$. As this is the “easiest” item, students did not have to show a very high level of ability to respond correctly to this item. Figure 12 shows the item information curve (IIC) as generated for Item 5 of the ISA. As almost everyone was correct on item 5, the IIC does not reveal very much information.

Item 2 is a polytomous item that accommodates the 0-6 scale’s response options. As seen in Figure 12 the ICC of item 2 provides a value for slope ($a=0.63$) and seven values for the threshold boundary between the seven response options ($d_0, d_1, d_2, d_3, d_4, d_5, d_6$, and d_7) (see Table 13). Item information in the GPCM (Generalized Partial Credit

Model) is calculated from the value of the slope parameter and the spread of the thresholds (Embretson & Reise, 2000), such that higher values for information have steeper slopes and the between-category threshold parameters for an item are distributed fairly evenly. Item 2 shows the highest value for information across all items, which implies that the item estimates ability levels more accurately than other items.

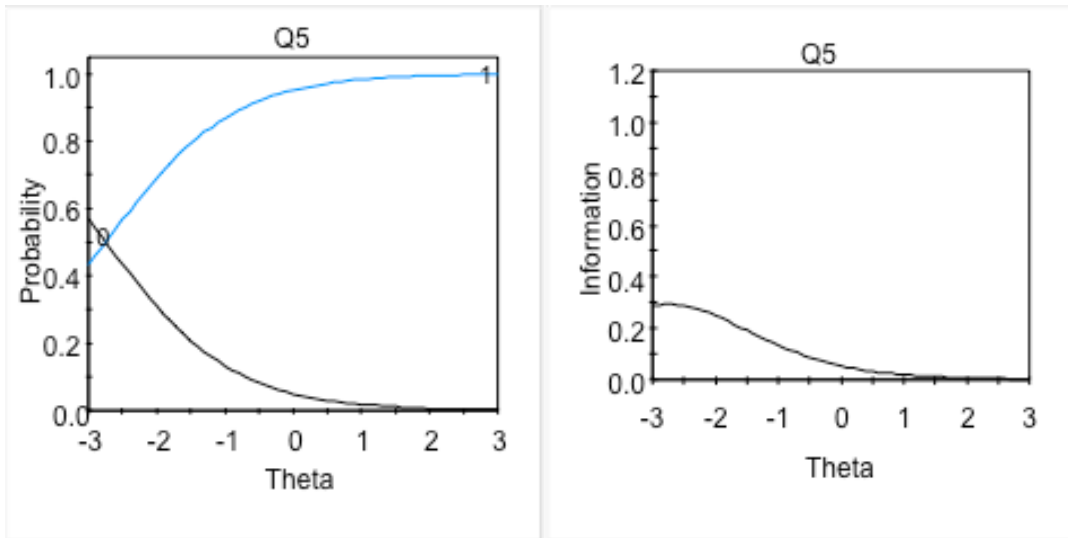


Figure 11. Item characteristic curve and item information curve for Item 5.

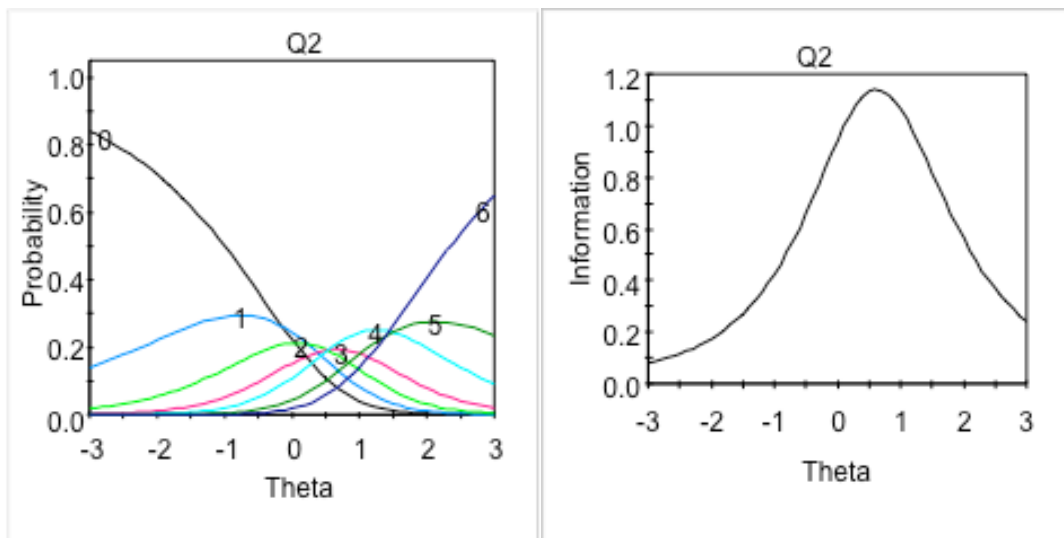


Figure 12. Item characteristic curve and item information curve for Item 2.

Facet map. Figure 13 shows a facet map for the ISA. The distribution of students' interdisciplinary understanding levels (indicated by * or dots) is displayed on the left side of the map whereas those on the right side are item difficulty values in logit scale (indicated by I1 for Item 1, etc.). The distribution of ISA subscales, interdisciplinary and disciplinary, is located in the middle of the map. Although the results support the unidimensional structure of the ISA, the facet map suggests that the two ISA subscales may exist, showing relatively different difficulty levels. In the map, the interdisciplinary items on average have higher difficulty levels than the disciplinary items. The figure shows that students' performance levels are distributed between -2 and 2 on the logit scale and the items within the subscales show an acceptable spread between -1 and 1 on the logit scale. Items were located at each point on the scale to measure meaningful differences but items did not cover all the areas on the ruler to measure the ability of all students. Thus, the ISA measure requires more items to cover students with the highest performance (between logit 1 and 2) and the lowest performance students (between -2 and -1). Importantly, the map indicates that item 5 is the easiest item and the item does not assess the students' interdisciplinary understanding level. This informs us that the item should be excluded from the ISA measure for a robust assessment tool.

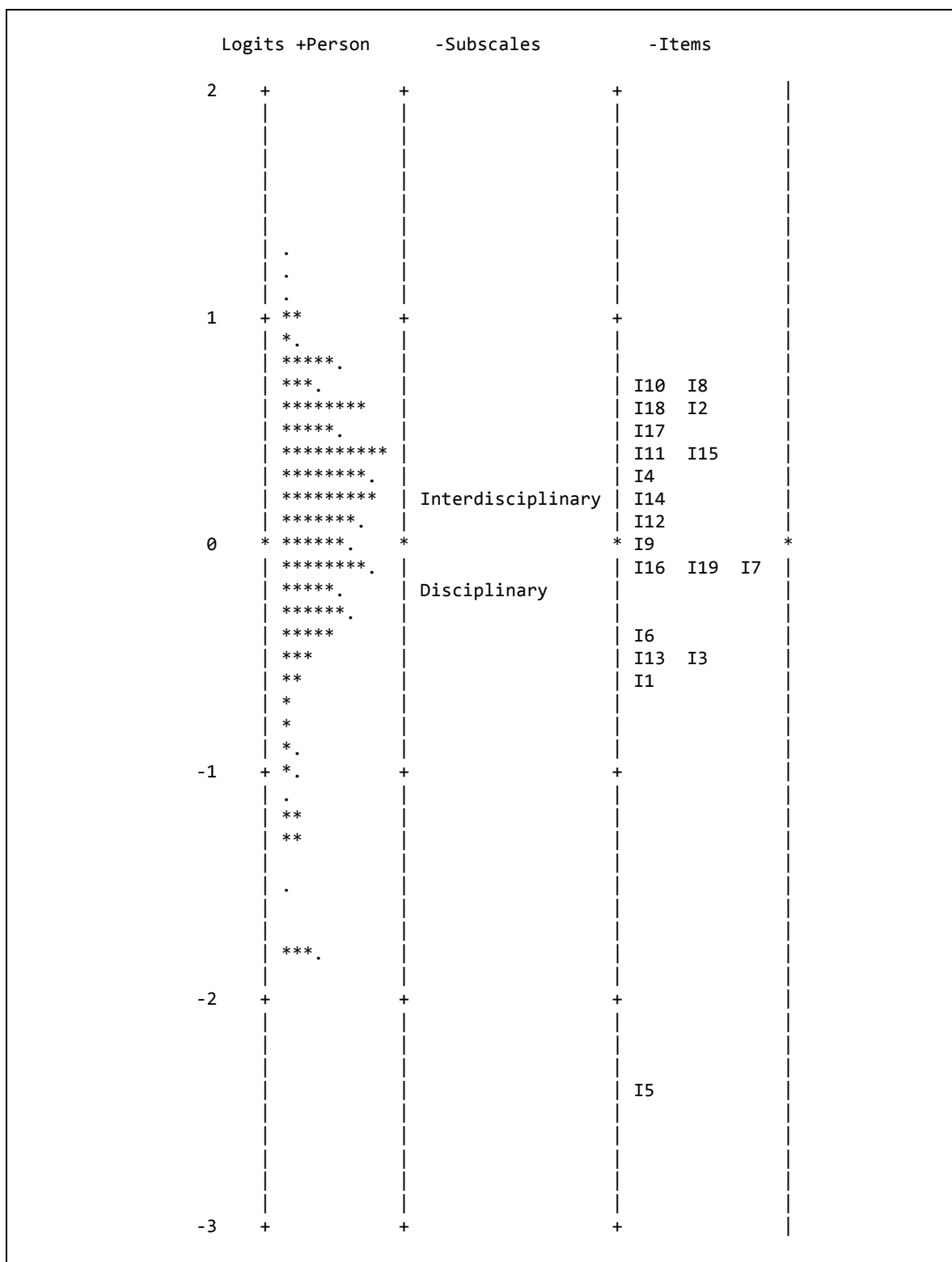


Figure 13. Facet map of the ISA.

Differential item functioning. In this study, DIF offered a means to identify gender-biased items. IRTPRO uses the Wald test (Lord, 1980) where DIF can occur when discrimination parameters and threshold parameters (or difficulty parameters) vary across sub-population. Through a chi-square method, the statistical significance of DIF across gender was evaluated. Table 14 presents the items with DIF across gender. Item 6 shows non-uniform DIF; the gender difference in item 6's difficulty changes across the ability continuum, and the discrimination parameter for this item varies across gender. Item 14 shows uniform DIF across gender, indicating that this item is systematically more difficult for female groups than for male groups with the same ability. The items with DIF may cause bias, which in turn may lead to a negative impact on the construct validity.

Table 14. DIF statistics for the ISA items (* DIF items at 0.05 level).

| Item | Total χ^2 | <i>d.f</i> | <i>p</i> |
|------|----------------|------------|----------|
| 1 | 1.2 | 2 | 0.544 |
| 2 | 4.5 | 7 | 0.726 |
| 3 | 1.6 | 2 | 0.451 |
| 4 | 14.1 | 7 | 0.050 |
| 5 | 1.0 | 2 | 0.612 |
| 6 | 8.6 | 2 | 0.014* |
| 7 | 10.9 | 9 | 0.286 |
| 8 | 9.0 | 9 | 0.443 |
| 9 | 1.0 | 2 | 0.609 |
| 10 | 2.8 | 2 | 0.248 |
| 11 | 2.0 | 2 | 0.361 |
| 12 | 4.1 | 9 | 0.902 |
| 13 | 1.7 | 2 | 0.424 |
| 14 | 17.6 | 9 | 0.041* |
| 15 | 7.3 | 9 | 0.606 |
| 16 | 4.9 | 2 | 0.084 |
| 17 | 2.6 | 2 | 0.268 |
| 18 | 0.1 | 2 | 0.946 |
| 19 | 15.3 | 9 | 0.082 |

Reliability. The reliability (internal consistencies) of the items was assessed by Cronbach's alpha. The internal consistency for all 19 items was 0.782. This shows modest reliability for the instrument, according to Nunnally & Bernstein (1994). To improve reliability, consideration was given to removing items but when items were deleted, the reliability was lower or the same; thus, all items were retained.

Hypothesized latent factor structures of the ISA. Table 15 is a display of Point-Biserial correlation coefficients among the 19 items of the ISA and two subscales. The items comprising each of the subjective sub-scales had a very wide range of correlations with their own sub-scales, from 0.219 to 0.858. All items were also significantly correlated with the other sub-scale, although some correlations were considerably lower. It was found that the correlation between interdisciplinary subscale and disciplinary subscale was high and significant ($r=0.701, p<0.001$).

Table 15. Point-Biserial correlations between individual items subscales.

| | Disciplinary | Interdisciplinary |
|-------------------|--------------|-------------------|
| Q1 (I) | .222** | .259** |
| Q2 (D) | .858** | .556** |
| Q3 (I) | .178** | .253** |
| Q4 (D) | .751** | .580** |
| Q5 (D) | .336** | .215** |
| Q6 (D) | .382** | .218** |
| Q7 (I) | .517** | .707** |
| Q8 (I) | .401** | .645** |
| Q9 (I) | .222** | .300** |
| Q10 (I) | .272** | .355** |
| Q11 (I) | .289** | .278** |
| Q12 (I) | .441** | .668** |
| Q13 (I) | .323** | .361** |
| Q14 (I) | .392** | .630** |
| Q15 (I) | .456** | .603** |
| Q16 (D) | .396** | .305** |
| Q17 (D) | .423** | .356** |
| Q18 (I) | .177** | .219** |
| Q19 (I) | .435** | .592** |
| Disciplinary | 1 | .701** |
| Interdisciplinary | .701** | 1 |

Note: ** $p < 0.001$

This study identified three plausible models to examine the latent factor structure of the ISA (see Table 16). One is an orthogonal two-factor model, which was used as a base line model. The orthogonal model was not satisfactory, indicating very poor fit. The oblique two-factor model considering the correlation between disciplinary and interdisciplinary factor produced a good fit. As a result, it is inferred that there is a strong correlation relationship between two factors. The third model is the second-order model. This model does not allow for two sub-factors to correlate, but rather their co-variance is

explained by the second-order construct. The second-order factor model implies that there is another latent construct that governs the first order factors. Both the second-order factor and the oblique two-factor model displayed the same acceptable levels of fit, but theoretical grounding supports the second-order factor model as the most plausible model for a good description of the ISA structure. Confirming the internal structure of the ISA hypothesized from the beginning step of item design provides evidence of the structural validity of the ISA.

Table 16. Summary of Fit Statistics for the Models Tested.

| | χ^2 | df | p | χ^2/df | CFI | TLI | RMSEA (90% CI) |
|-------------------------------|----------|------|--------|-------------|-------|-------|---------------------|
| Two-factor model (orthogonal) | 1175.748 | 152 | <0.001 | 7.735 | 0.248 | 0.154 | 0.122 (0.115:0.128) |
| Two-factor model (oblique) | 243.595 | 151 | <0.001 | 1.613 | 0.932 | 0.923 | 0.037 (0.028:0.045) |
| Second-Order model | 243.595 | 151 | <0.001 | 1.613 | 0.932 | 0.923 | 0.037 (0.028:0.045) |

The relationship between the number of science courses taken and ISA

performance. The Pearson correlation between the ISA score and the number of science courses taken shows a weak but statistically significant relationship ($r = .275, p < 0.001$). In order to compare the ISA score means, I categorized students into four groups. Groups 1, 2, and 3 represent students who have taken one-three, four, and five science courses, respectively. Group 4 is students who have taken six and more science courses both in high school and college. The descriptions of the four groups and frequency distribution are found in Table 17.

Table 17. Descriptive statistics of four groups categorized by the number of course taken.

| Groups | N | Mean (θ) | SD (θ) | Min (θ) | Max (θ) |
|--------|-----|-------------------|-----------------|------------------|------------------|
| 1 | 152 | -0.3661 | 0.9083 | -2.51 | 1.60 |
| 2 | 77 | -0.1079 | 0.9450 | -2.19 | 2.03 |
| 3 | 69 | 0.0918 | 0.8389 | -2.17 | 1.59 |
| 4 | 152 | 0.3686 | 0.8002 | -1.91 | 1.89 |
| Total | 454 | -0.0003 | 0.9190 | -2.51 | 2.03 |

Note: 1: students who have taken only one science course so far, 2: students who have taken two science courses so far, 3: students who have taken three science courses, 4: students who have taken more than 4 science courses.

This study used theta values of ISA performance provided by IRT as a dependent variable, not the summed total score on the ISA. The assumptions of normality ($p=0.121$ in Kolmogorov-Smirnov test) and homogeneity of variance ($p=0.099$) were satisfied. Thus, a one-way ANOVA and Tukey post-hoc tests were carried out and there was a statistically significant difference among four groups, $F(3, 450) = 19.015, p < 0.001$. Tukey post-hoc tests indicated that students who have taken more than six science courses showed a significant difference when compared with group 1 and 2 students. Also, there is a significant difference in interdisciplinary performance between students who have taken five courses and students who have taken one-three. There is no statistically significant difference between group 1 vs. 2, 2 vs. 4, and 3 vs. 4.

The relationship between the grade levels and ISA performance. The grade levels were divided into five groups (high school students, college freshman, sophomore, junior, and senior/graduates). The equal variance assumption was satisfied among five groups ($p=0.288$). The five groups' mean distributions are shown in Table 18. Performance on the ISA differed significantly across grade levels, $F(4, 449)=32.483, p < 0.001$. Post-hoc tests revealed that college students performed much better than high school students (p

<0.001) in interdisciplinary understanding. No statistically significant differences were detected among college students.

Table 18. Descriptive statistics for grades.

| Grade levels | N | Mean (θ) | SD (θ) | Min (θ) | Max (θ) |
|-----------------|-----|-------------------|-----------------|------------------|------------------|
| High school | 44 | -1.3108 | 0.7363 | -2.17 | 1.08 |
| Freshman | 186 | 0.1043 | 0.7955 | -2.51 | 1.89 |
| Sophomore | 164 | 0.1224 | 0.7960 | -2.07 | 2.03 |
| Junior | 39 | 0.3565 | 0.9614 | -2.01 | 1.88 |
| Senior+Graduate | 21 | 0.1998 | 0.9476 | -1.91 | 1.69 |
| Total | 454 | -0.0003 | 0.9190 | -2.51 | 2.03 |

The relationship between race and ISA performance. Table 17 shows the distribution of the ISA means across eight race groups (group 2's N is 0). Only three race groups, White, Asian, and Hispanic, were selected for ANOVA because other groups did not have a large enough sample size to conduct ANOVA. The assumption of normality was met but equal variance was not assumed; thus, two alternative F tests, Welch and Brown-Forsythe, were performed. Both the F statistics showed a statistically significant difference among three groups, indicating $F(2, 239.79) = 11.272, p < 0.001$ in Welch statistic and $F(2, 328.094) = 12.722, p < 0.001$ in Brown-Forsythe statistic. The Games-Howell post-hoc test is appropriate when the equal variances assumption has been violated. This test revealed that significant differences between White and Hispanic, and between Asian and Hispanic students.

Table 19. Descriptive statistics of ISA theta scores for race.

| Race | N | Mean (θ) | SD (θ) | Min (θ) | Max (θ) |
|--|-----|-------------------|-----------------|------------------|------------------|
| White | 177 | 0.1026 | 0.8644 | -2.51 | 1.85 |
| Asian | 131 | 0.2185 | 0.8303 | -2.19 | 1.88 |
| Native Hawaiian or Pacific Islander | 2 | -0.5070 | 0.3875 | -0.78 | -0.23 |
| African American | 23 | -0.2622 | 0.8363 | -2.1 | 0.97 |
| Hispanic | 106 | -0.3578 | 1.0309 | -2.17 | 2.03 |
| Other | 15 | -0.1291 | 0.8581 | -1.54 | 1.89 |
| Total | 454 | -0.0003 | 0.9190 | -2.51 | 2.03 |

Note: The sample size of group 2 (American Indian or Alaska Native) is 0

The relationship between gender and ISA performance. An independent t-test was conducted with gender as a grouping variable. There was no significant difference in mean score in the ISA between female students and male students, $t(452) = 0.053$ ($p = 0.958$). This result indicates gender equality in the outcomes of the ISA.

Chapter 5: Discussion and Conclusions

This chapter discusses the implications of the interdisciplinary science assessment through a comparison with the existing research literature. This chapter also summarizes the main findings by focusing on construct validity for the instrument tool, and describes implications for the use of the ISA for instructional purposes, directions for future research, limitations, and conclusions.

5.1. Implications of the interdisciplinary science assessment through a comparison with other literature

Even though there is ample argument for the significance of the interdisciplinary approach both in terms of science learning and targeted education outcomes, to date, there have been few empirical studies assessing students' interdisciplinary understanding in the area of K-16 science education (e.g., Shen, Liu, & Sung, 2014). Shen, Liu, and Sung (2014) reported on the development of a tool for assessing college students' understanding of a single topic (osmosis), a topic that involves knowledge from multiple science disciplines. They identified the key concepts related to osmosis through content experts' group discussion and their concept maps. Through the item refining process, the authors included both 15 disciplinary items and 25 interdisciplinary ones in their final version of the survey. What Shen et al.'s study and the current study have in common is that both seek to develop a set of items targeting students' interdisciplinary understanding in a given topic, which requires students to integrate different sets of key concepts from different science disciplines. However, there are several differences between these studies in the item development process. The current study adopted the construct modeling

framework to develop the items through a more systematic process and used a different IRT model and factor analysis to establish a more robust construct validity. Also, the current study detailed the rubric development process including the inter-rater reliability issue for CR items based on a systematic evaluation process.

There is a small body of literature reporting the development of assessment tools using the framework of knowledge integration (e.g., Liu et al., 2008; Lee & Liu, 2010), which is similar to interdisciplinary understanding. Liu et al. (2008) and Lee and Liu (2010) adopted items from existing pools of standardized tests such as NAEP and TIMSS, but did not develop new items that required interdisciplinary understanding. All their items contained diverse contents from different science disciplines, but individual items still emphasized knowledge from each single discipline rather than focusing on integrated knowledge or ideas from multiple science disciplines.

Similarly, Zwickle et al. (2014) developed an assessment to measure sustainability knowledge. The items were developed in three domains: environmental, economic, and social, with the help of several experts across diverse academic disciplines such as ecology, sociology, education, etc. According to the definition of ‘interdisciplinarity’ defined in the chapter 2 of the present study, the developed instrument followed a multidisciplinary approach rather than an interdisciplinary one. ‘Multidisciplinary’ requires two or more disciplines that do not necessitate the integration of knowledge (Chynoweth, 2009; Klein & Newell, 1997). ‘Multidisciplinarity’ does not have connections at all between disciplines involved and indicates juxtaposition of various disciplines with no apparent connection between them (e.g. environmental +

economic + social) (Kochelmans, 1979).

Lack of sufficient interdisciplinary assessment tools in the area of science education compels the need for a new instrument to be developed in order to address existing challenges for real interdisciplinary learning. This study contributes to the field through development of a new instrument that captures students' interdisciplinary understanding of the carbon cycle. The topic of the carbon cycle is intrinsically interdisciplinary with combined ideas and information from different disciplines (de Baar & Suess, 1993). Carbon can be found with different forms among the biosphere, pedosphere, geosphere, hydrosphere, and atmosphere of the Earth. Moreover, the carbon cycle is related to our civilizations, economies, and history (e.g., the industrial revolution). The movement of carbon in the Earth system and the interaction between natural systems and human activities allow for a rich interdisciplinary context, which enables the creation of interdisciplinary items.

5.2. Addressing construct validity

Ensuring that an assessment measures what it is intended to measure is a critical component in test development. Orpwood (2007) argued that there is a lack of validated tests and assessments available in science education. This suggests that there is a need for psychometrically sound work in the field of science education. The CFA and IRT analyses in this study yield detailed insights into the psychometrics of the ISA to establish construct validity. The findings suggest that overall the ISA showed satisfactory psychometric quality in measuring students' interdisciplinary understanding of carbon cycling even though there is room to improve the scale through the deletion of one misfit

item. The fit statistics in the IRT models are used to check the relevance of the intended test construct. Misfitting items show a possibly different and irrelevant construct. Misfit can arise from diverse reasons. In this study only one misfitting item (i.e., item 5) was found. Item 5 is a MC question asking the student to fill in the two blanks in the excerpt provided with a common gas, with the correct answer being CO₂. In the Wright map comparing the relative difficulty of items and the ability spread of students (see Figure 13), the ability level (theta value) of all students is higher than item 5's difficulty level, which allows us to expect that all students should get the item right. However, 7.7% of the students answered this item incorrectly. Thus, the unexpected observation of students with a higher ability level than item 5's difficulty level getting the wrong answer generated a misfit item (Boone, Staber, & Yale, 2014). The Wright map also allows for verification of the representativeness of a test construct by identifying the extent of item coverage by the amount of redundancy and the range of the interdisciplinary understanding of carbon cycling in the sample (Smith, 2001). The items were written based on a construct map informed by a hypothesis positing both interdisciplinary and disciplinary understanding. The item map was used as evidence congruent with construct validity about the inclusion of disciplinary and interdisciplinary items.

Confirming two sub-constructs using the theory-driven approach is relevant to structural validity because appropriate interpretation of scale scores can only be achieved by validation of the internal structure of the instrument. Among the proposed three internal structures including both domains (interdisciplinary and disciplinary), the second-order factor model is the best model to explain the ISA, having excellent model

fit indices. The second-order factor model supports the theoretical hypothesis of this research: the carbon cycling (the second order factor) is directly associated with two underlying sub-constructs (interdisciplinary and disciplinary) and each sub-construct is measured using a certain number of items.

5.3. Making inferences about interdisciplinary understanding scores

Score interpretation made based on scores is an essential consideration in construct validity (Messick, 1988). The descriptive and inferential statistics help interpret the scores. Students showed limited interdisciplinary understanding with the mean score of 28.57 out of 71. As grade level increased, the interdisciplinary understanding performance increased even though there was not a clear positive linear relationship between performance and grade. A t-test supported the fact that interdisciplinary understanding of college students (mean = 30.29) is much higher than high school students, (mean = 11.39) showing a statistically significant difference.

In terms of the number of science courses taken, an interesting result is that students who have taken more than six science courses in both high school and college had higher performance than students who took one to three or four courses. This result supports the fact that taking more than six science courses or advanced courses helps students' interdisciplinary learning: if students do not have such opportunities, they may experience a lack of integration of relevant knowledge.

5.4. Cognitive process between disciplinary and interdisciplinary learning

In terms of how students develop their interdisciplinary understanding, this study hypothesized that they learn given science content starting from knowledge of one

discipline through subsequent cognitive levels of more integrated understanding of the topic in a developmental view of learning. Students' initial knowledge or ideas based on unidisciplinarity are progressively integrated and elaborated towards a more desirable interdisciplinary understanding. The construct map was developed as shown in the previous chapter, describing the nature of development in interdisciplinary learning and thus serves as a frame of reference for monitoring individual growth. In this perspective, I intended to develop both disciplinary and interdisciplinary items requiring different levels of interdisciplinary understanding. Interdisciplinary items refer to items that require knowledge from more than one discipline on the part of students to answer them perfectly. Disciplinary items, on the other hand, call for knowledge of only one science discipline.

Zhang and Crawford (2014) proposed the three levels of cognitive process for learning crosscutting concepts. As shown in Figure 14, they asserted that 'coalescence' needs to occur when learning crosscutting concepts from the disciplinary level (level 2) to the interdisciplinary level (level 3), in which students' understanding of crosscutting concepts moves beyond the limitation of specific scientific disciplines and students can understand crosscutting concepts in a more interdisciplinary manner. Their model supports the current study's argument that disciplinary learning from multiple disciplines is prerequisite for interdisciplinary learning.

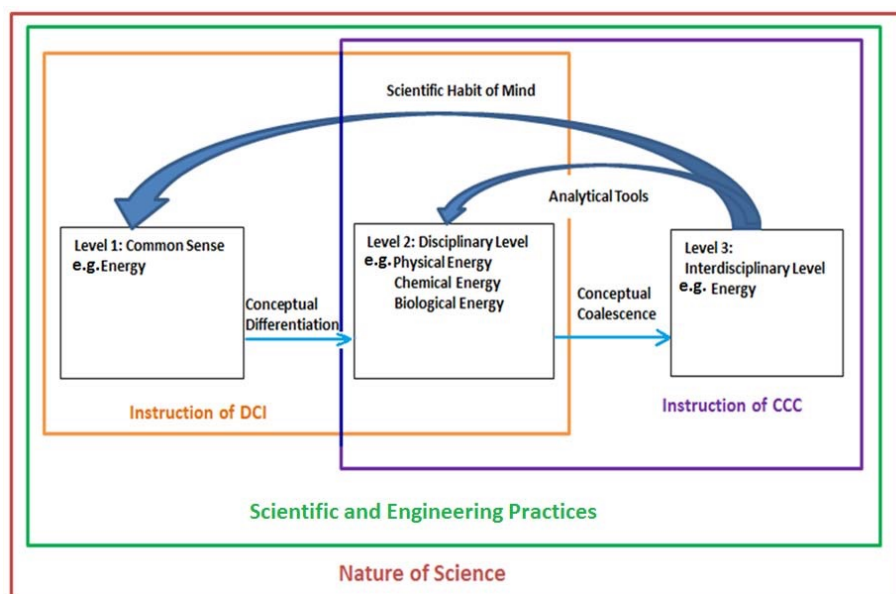


Figure 14. Model of cognitive process of crosscutting concepts in NGSS (DCI: Disciplinary Core Ideas, CCC: Crosscutting concept). Reprinted from “Learning and Teaching Crosscutting Concepts from Cognitive Perspectives,” by Zhang and Crawford (2014, p.3), Reprinted with permission.

5.5. Linking interdisciplinary science assessment to instruction

The development of interdisciplinary assessment has effects on teachers’ choice of instructional strategies. Wiggins and McTighe (1998) invented “Backward Design” to make the influence of assessed outcomes on the design of instruction explicit. This method is backward to traditional curriculum planning. Backward design begins with teachers thinking about assessment before deciding what and how they will teach as a purposeful task. While teachers start creating a list of content that will be taught in backward design curriculum planning, they first clarify goals and create assessments, after which they craft lesson plans to achieve those goals in. The development of the precedent interdisciplinary assessment enables teachers to design their instructional practices in a way that students can learn how to connect one concept to another across

different science disciplines in order to construct explanations and form arguments rather than placing emphasis on memorizing facts. Such instructional environments should eventually lead to an excellent interdisciplinary understanding of the given interdisciplinary issues and desirable scientific literacy.

The shift in the roles of assessment by focusing on interdisciplinary understanding as a new learning outcome is associated with a variety of changes in teachers' instructional practices. It is thus important to understand how the adoption of a new and reformed assessment would likely influence teachers' and students' activities in classrooms. The specific purposes for which an assessment will be used need to be taken into critical consideration for its future use. For example, assessments can be used to demonstrate students' understanding and inform instructional decisions by instructors in classrooms, to compare students' achievement among various schools, or to ensure accountability for the teachers. The ISA was designed as an internal classroom assessment. According to Ruiz-Primo, Shavelson, Hamilton, and Klein (2002), assessments are divided into five levels (i.e., immediate, close, proximal, distal, and remote) based on the purpose of the assessment. Adopting their framework, the ISA could be understood as either a 'close' or 'proximal' level on this continuum because it is assumed that carbon cycling knowledge and interdisciplinary understanding are sensitive to the content and activities of the curriculum. The ISA will thus be used for semiformal tests of learning from one or more lesson units or formal classroom exams of learning using specific instructional practices in order to recognize and respond to students' interdisciplinary learning or to examine the effectiveness of the lessons, respectively.

5.6. Directions for future research

One direction for future research is continued improvement of the instrument. According to DeVellis (2003), the development of a new scale requires taking time to thoroughly investigate all aspects including theoretical grounding, multiple perspectives regarding the appropriateness of items included, participant pool, and statistical techniques. Because “construct validity is a never-ending, ongoing, complex process over a series of studies in a number of different ways” (Pett et al., 2003, p.29), the addition of new items or the deletion of the items in the current assessment tool might lead to achievement of a more robust construct validity. If samples over 1000 are obtained, we could choose other IRT models for the analysis, for example, a three-parameter logistic model (3-PLM), which even examines the guessing parameter for MC items to confirm further validation.

The second direction for future research is assessing test fairness. This can be conducted by a statistical analysis of the psychometrics or interpretation of test scores depending on group membership such as gender or race. If an assessment item is relatively more difficult for members of one group than for members of another after controlling for the overall levels of the groups on the construct of interest, then it will need a revisit (Reynolds & Kaiser, 1990). The differential item functioning (DIF) in this context it implies that different groups have different probabilities of getting an item correct. This study tested the DIF for gender but further fairness testing is needed for race and other group characteristics. Beyond the IRT method used here, structural equation

modeling techniques with multiple indicators and group comparison methods could be used to evaluate the hypotheses of measurement invariance over different groups.

The third direction for future research is to examine how students' interdisciplinary understanding changes during an interdisciplinary science course developed based on the framework of backward design. Also, through detecting the changes in students' interdisciplinary understanding performance, empirical evidence supporting the rationale for the effectiveness of instructional practices aimed at interdisciplinary learning can be provided. Longitudinal studies or studies with quasi-experimental settings are required to address the research topics above.

The fourth direction for future research is to investigate cultural validity as a form of test validity in assessment. It is assumed that cultural background influences the ways in which students interpret items and the cognitive processes they use in completing those items. To establish cultural validity, the items can be administered to a sample of students from other countries and the U.S. This cultural validity allows us to determine 1) whether students from different cultural groups exhibit different patterns in assessment; 2) how culture influences the cognitive process of interdisciplinary understanding; and (3) whether those differences can account for performance score differences among various cultural groups.

5.7. Limitations

This study posited that interdisciplinary understanding can be a more advanced cognitive domain than unidisciplinary understanding. However, an opposing hypothetical framework may also have great potential in informing us about how both domains

interact with each other. Thus, constructing the framework by empirical evidence would be worthy of future investigation.

Another limitation to the study is the imbalanced sample size of college and high school students. The small sample size of high school students in comparison with college students is less representative of the population, and there is some possibility of creating an undue influence for outliers or extreme observations. Thus, if this study had a greater sample size of high school students, there would be the possibility of having more robust statistical results.

A final limitation is that this study tried to investigate DIF for race but the software used (IRTPRO) did not provide an estimation of the DIF. In cases where there is no value assigned in the scale range, the program will not function properly. Thus, other programs (e.g., Mplus (Muthén & Muthén, 1998-2015) or Winsteps (Linacre, 2015)) with other estimation processes will be used for future research.

5.8. Conclusions

As the education paradigm moves from being disciplinarily focused to interdisciplinarily focused, the science education community needs to reflect and evaluate exactly what skills and abilities all students should be equipped with in the 21st century. This study argues that interdisciplinary understanding is one of the overarching components that students should develop when learning science, mainly because it enables students to explain the natural world and to make decisions that could impact the future of that world through human activities with scientific knowledge from different disciplines. Interdisciplinary understanding can be obtained from interdisciplinary

learning experiences where students start from recognizing the connections among scientific concepts, having a holistic view on a given phenomenon. In order to measure this interdisciplinary understanding of students in the field of science and STEM education, this study presents a new instrument for measuring interdisciplinary understanding in science that cannot be directly assessed with traditional tests. I propose that this can be a useful instrument for the field particularly in light of the prevailing calls for science education reform and the promotion of STEM learning. The results of the study provide encouragement for additional research that requires the development and validation of assessment instruments. Further, the systematic development and analysis processes used here can be expected to yield assessment tools that have strong psychometric properties and will be valuable for teachers in the classroom.

Appendixes

Appendix A: Interdisciplinary science assessment (ISA)

Interdisciplinary Science Assessment (ISA) consist of:

- Questions related to the global carbon cycle (Section A)
- Questions about yourself (Section B)

Instructions

- Please read each question carefully and answer as accurately as you can
- In the open-ended questions, **use all the concepts and knowledge that you have learned so far in science classes** (e.g., earth science, chemistry, physics, biology etc.) to write your answer for each question.
- For section B, you will normally answer by checking a box.

SECTION A: INTERDISCIPLINARY UNDERSTANDING

Open ended questions – Write down all possible answers

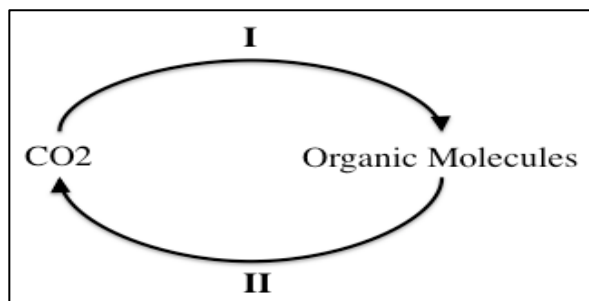
Multiple choice questions – Tick (✓) the one correct answer

Answer all questions

Item 1) Carbon cycling describes the movement of carbon (typically bound with other elements in compounds) through Earth's atmosphere, hydrosphere (oceans and other bodies of water), biosphere (plants and animals), and lithosphere (rocks and soils). Carbon exists in various chemical forms in each sphere. Which of the following carbon forms is NOT found in the sphere indicated?

- A) Carbon dioxide in the atmosphere
- B) Calcium carbonate in the lithosphere
- C) Glucose in the biosphere
- D) Bicarbonate and carbonate in the hydrosphere
- E) None of the above; all ARE found in the spheres indicated.**

Item 2) Carbon continuously cycles through an ecosystem. A simplified carbon cycle showing step I, where organic molecules are created, and step II, where CO₂ is produced, is depicted below.



Describe the role of Step I and Step II in carbon cycling respectively, and explain how each of the processes in Step I and Step II promotes the movement of carbon within the carbon cycle.

Item 3) The most popular timber product grown in the United States today is *Pinus taeda* known as Loblolly Pine. The pine trees' average height and circumference are reported to have been increasing since the 1960s. The level of CO₂ in the atmosphere has also been increasing rapidly since 1950. It has been proposed that the increased burning of fossil fuels might explain the increased growth for this species. Among the following statements, which one do you agree with regarding this proposed explanation?

A) This explanation would be difficult to test because of all the other factors, such as temperature and light level, which might have affected the growth of the trees.

B) This explanation makes sense because an increase in atmospheric CO₂, one of the inputs in photosynthesis, will always lead to increased glucose production and more plant growth.

C) This explanation cannot be right because increased CO₂ causes global warming, which is detrimental to plant growth.

D) This explanation cannot be right because the burning of fossil fuels releases sulfur dioxide (SO₂), which causes acid rain and kills plants.

E) None of the above

Item 4) The paragraph below describes a food chain on the interrelationship that exist between organisms in a field ecosystem. After reading this paragraph, answer the following question.

Oak trees produce large autumnal acorn crops every two to five years. Acorns are a critical food source for the white-tailed deer. Coyotes are important predators of the white-tailed deer. Decomposers consume the remains of dead white-tailed deer, coyotes, and acorns and then return some of nutrients from the dead organic matter back into the soil.

Describe how carbon produced from oak trees are transferred within the food chain through consumers to decomposers and also show how energy is transferred through the food chain.

Item 5) When animals feed on green plants, they pass carbon compounds to other animals in the upper levels of the food chain. Animals release _____ into the atmosphere

during respiration. _____ is also released when plants and animals die. This occurs when decomposers (bacteria and fungi) break down dead plants and animals (decomposition) and release the carbon compounds stored in them. Which word is appropriate for both blanks?

- A) Carbon dioxide
- B) Oxygen
- C) Nitrogen
- D) Argon
- E) Methane

Item 6) Decomposers, such as fungi and microbes, will affect the carbon cycle by...

- A) consuming carbon, reducing the total amount of carbon that exists in the cycle
- B) releasing CO₂ into the atmosphere, increasing atmospheric levels of carbon**
- C) creating carbon from other atomic elements (such as oxygen)
- D) hindering producers from getting the nutrients they need
- E) None of the above; decomposers do not affect the carbon cycle

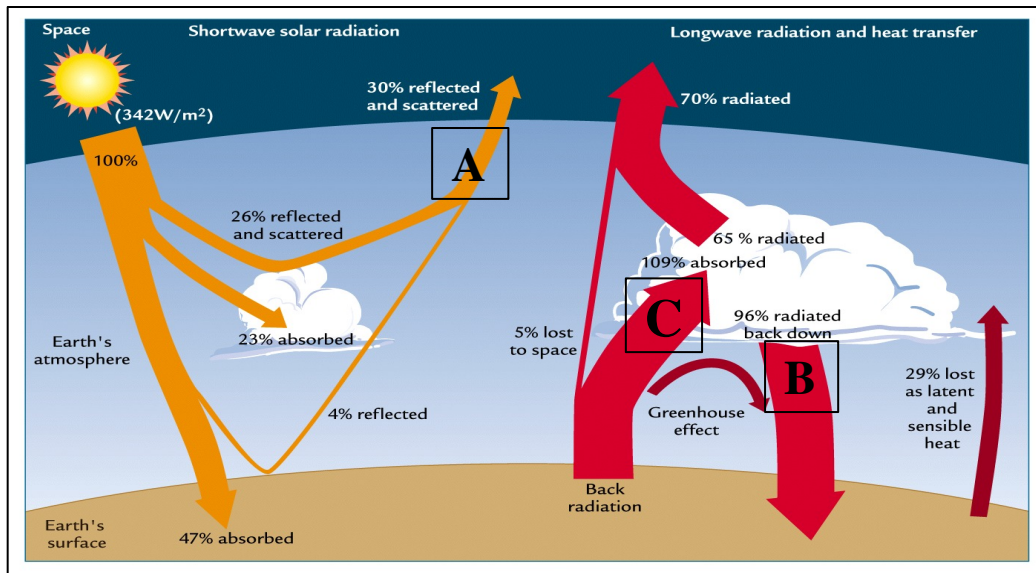
Item 7) Fossil fuels such as coal, oil and natural gas contain high concentrations of hydrocarbons from the organic remains of prehistoric organisms. The following chemical equation represents the combustion reaction of natural gas (Natural gas is mainly methane). Complete the chemical equation.



b) How is the equation similar to or different from the equation for cellular respiration?

Item 8) Scientists have recently calculated that approximately 26% of all CO₂ emitted from human-related activities, such as the combustion of fossil fuels, was absorbed by oceans during the decade 2002-2012. This resulted in 2.5 billion gigatons of excess carbon moved from the atmosphere into the ocean each year over the course of a decade. Scientists are concerned that the mass of CaCO₃ deposited annually in coral reefs is decreasing. They expect that in 2050 the total amount of CaCO₃ in coral reefs will be 20 percent less than it is currently.

Use this information to describe how the combustion of fossil fuels might affect the loss of coral reefs in ocean water, even though the amount of CO₂ in the atmosphere is increasing.



Item 9) The figure above shows Earth's energy budget diagram, where A, B and C label energy transfer processes. Which of the following statements is NOT true?

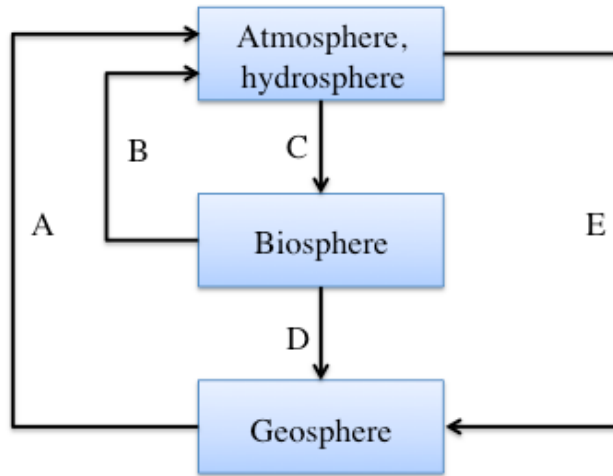
- A) If polar ice sheets melt due to global warming, A will increase.
- B) As B increases, global warming will increase.
- C) If the temperature of the Earth's surface increases, C will also increase.
- D) When the concentration of the main greenhouse gases rises, B and C will increase.
- E) Long wavelength (red arrows) radiation is absorbed by greenhouse gases, which in turn re-radiate much of their energy to the surface and lower atmosphere.

Item 10) The Earth has a natural mechanism that prevents it from overheating on long (hundred thousand year) timescales. Which process completes the **negative feedback** relationship that describes this mechanism?

Atmospheric CO_2 concentrations increase \rightarrow global mean temperature increases \rightarrow the weathering rate of silicate mineral increases \rightarrow _____

- A) the rate of physical weathering increases
- B) plants grow faster
- C) plants grow slower
- D) atmospheric CO_2 concentrations increase
- E) **atmospheric CO_2 concentrations decrease**

Item 11) The figure below represents the cycling of carbon through the Earth's spheres. Which of the following statements about the role of carbon dioxide (CO₂) in the carbon cycle are correct?

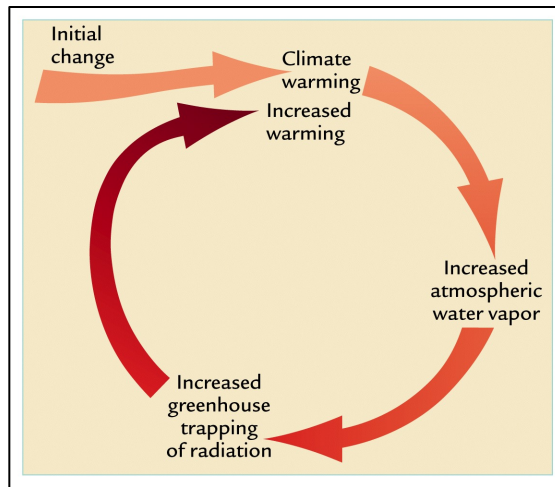


- I. In process A, a huge amount of CO₂ is emitted through volcanic eruptions.
- II. In C, solar energy is converted into chemical energy
- III. B represents the respiration of organisms
- IV. D and E represent the formation of limestone

- A) II only
- B) I, II
- C) II, III
- D) II, III, IV
- E) I, II, III, IV

Item 12) The figure below shows positive feedback cycle. In the context of global warming,

- a) What could the initial change be in this process?
- b) Discuss how the initial change leads to the feedback loop in the diagram below.

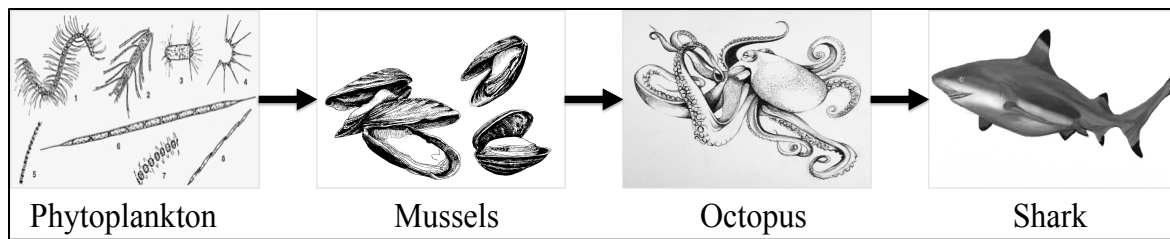


Item 13) Which of the following statements about the role of carbon dioxide (CO₂) in the carbon cycle are correct?

- I. Carbon dioxide is produced during photosynthesis
- II. Carbon dioxide concentration in the atmosphere increases when trees are cut down and when trees decay
- III. The primary non-anthropogenic source of carbon dioxide in the atmosphere is carbon dioxide being released from the Earth's oceans.

- A) I only
- B) II only
- C) III only**
- D) II and III
- E) I, II and III

Item 14) A Figure below shows one example of a food chain in an ocean ecosystem. The emission of anthropogenic CO₂ into the atmosphere causes changes to ocean ecosystems. How could the increased amount of CO₂ in the atmosphere affect the food chain of the ocean ecosystem?



Item15) The geological carbon cycle is complicated, with many different pieces playing their roles. Usually a change in one part of the cycle causes compensating changes in other parts, but sometimes the system takes a long time to get back into balance. For example, it takes a long time for the oceans to increase their uptake of carbon dioxide, so they might not be able to compensate for CO₂ increase in the air, resulting in an imbalance. How could deforestation lead to an imbalance in carbon dioxide levels?

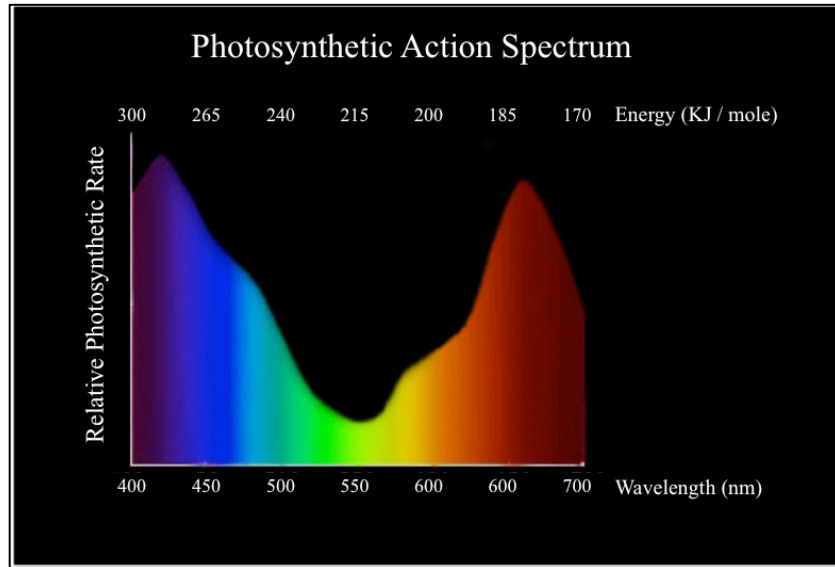
Item 16) If ice on the Earth's surface melts, it affects the Earth's albedo (i.e., the percentage of sunlight that the Earth reflects back into space). Which of the following statements would be true regarding the effect of melting ice and its effect on the Earth's albedo?

- A) Melting ice would increase the Earth's albedo resulting in less heat absorbed, which in turn would result in more ice melting. This would be known as a positive feedback loop.
- B) Melting ice would decrease the Earth's albedo resulting in more heat being absorbed, resulting in more ice melting. This would be known as a positive feedback loop.**
- C) Melting ice would increase the Earth's albedo resulting in more heat being absorbed, resulting in more ice melting. This would be known as a positive feedback loop.
- D) Melting ice would increase the Earth's albedo resulting in more heat absorbed, which in turn would result in less ice melting. This would be known as a negative feedback loop.
- E) Melting ice would decrease the Earth's albedo resulting in less heat being absorbed, resulting in less ice melting. This would be known as a negative feedback loop.

Item 17) Ocean acidification is caused when the ocean absorbs more CO₂ resulting in:

- A) A decrease in the pH of ocean water**
- B) An increase in the pH of ocean water
- C) A decrease in the temperature of ocean water

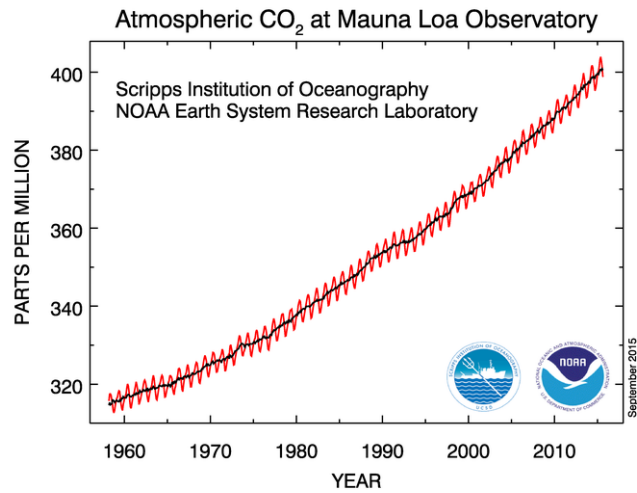
- D) An increase in the temperature of ocean water
- E) An increase in methane emissions



Item 18) The above figure shows the relationship between the electromagnetic spectrum of sunlight reaching the Earth and the relative photosynthetic rate. Photosynthesis depends upon the absorption of light by pigments in the leaves of plants. The most important of these pigments is chlorophyll. Which of the following statements is **NOT** true?

- A) Quantity of energy per photon is inversely proportional to the wavelength of the radiation.
- B) High energy photons are more likely to produce blue light rather than red light.
- C) Sunlight consists of many wavelengths of light, but only photons from certain portions of visible light can be taken up by chlorophyll.
- D) Green plants reflect green light and absorb red and blue light.
- E) **Chlorophyll pigments in plants can absorb the full range of wavelengths.**

Item 19) The graph below shows measurements of carbon dioxide in the atmosphere (measured in parts per million) taken at a NOAA station on Hawaii between 1958 and 2013. The **red line** represents a shorter-time scale variation of the monthly mean values of CO₂ in a given year. The **black line** represents a long-term trend over the last decades, after correction for the average seasonal cycle.



Some people claim that the fluctuations in the atmospheric CO₂ level are due to variations in radiation coming from the Sun.

- a) Describe how the amount of incident sunlight might affect the amount of carbon dioxide in the atmosphere.
- b) Are both of the **red line** and the **black line** pattern equally likely to be due to solar variations? Explain why or why not.

SECTION B: ABOUT YOU

Q1) What is your current education classification?

- ☐ Grade 9
- ☐ Grade 10
- ☐ Grade 11
- ☐ Grade 12
- ☐ College Freshman
- ☐ College Sophomore
- ☐ College Junior
- ☐ College Senior
- ☐ Graduate Students

Q2) Are you female or male?

Female ☐ Male ☐

Q3) Which of the following best describes your ethnic/racial background?

(Please tick only one box below.)

- ☐ White
- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Native Hawaiian or Other Pacific Islander
- ☐ Black or African American
- ☐ Hispanic
- ☐ Other (please specify) _____

Q4) What science courses have you taken so far? (Please check all the courses you have taken.)

- ☐ Physical Science
- ☐ Earth Science

- ☐ Chemistry
- ☐ Biology
- ☐ General Science
- ☐ Integrated Physics & Chemistry (IPC)
- ☐ Pre-AP Physical Science
- ☐ Pre-AP Earth Science
- ☐ Pre-AP Chemistry
- ☐ Pre-AP Biology
- ☐ Pre-AP Environmental Science
- ☐ Honors Physics
- ☐ Honors Earth Science
- ☐ Honors Chemistry
- ☐ Honors Biology
- ☐ AP Physical Science
- ☐ AP Earth Science
- ☐ AP Chemistry
- ☐ AP Biology
- ☐ AP Environmental Science

Others (please specify) _____

Thank you very much for your participation in this research!

Appendix B: Scoring rubrics

ANALYTIC SCORING GUIDELINES OF INTERDISCIPLINARY SCIENCE

ASSESSMENT

Table 1

Holistic rubric

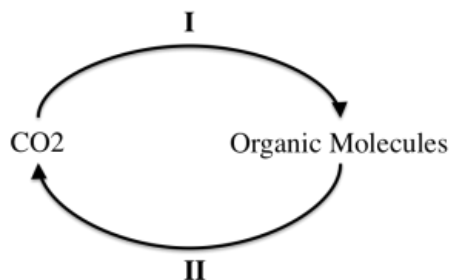
| Correctness | Interdisciplinarity | Description |
|---|---------------------|---|
| Partially/Fully correct | Interdisciplinary | A student uses relevant scientific concepts and principles to explain a specific event in carbon cycling to demonstrate his/her reasoned interdisciplinary understanding. |
| | Unidisciplinary | A student response has a partially or fully accurate understanding of the scientific concepts and principles from one necessary discipline to explain a specific event in carbon cycling. |
| Incorrect/Off-topic/Blank/Restatement of the prompt | No disciplinary | A student response is incorrect/not relevant/blank |

Table 2

Scoring scales

| Types of question | Perfect Score |
|-------------------|---------------------------------------|
| Disciplinary | 6 points (Q2, Q4) |
| Interdisciplinary | 8 points (Q7, Q8, Q12, Q14, Q15, Q19) |

Item 2) Carbon continuously cycles through an ecosystem. A simplified carbon cycle showing step I, where organic molecules are created, and step II, where CO₂ is produced, is depicted below.



Describe the role of Step I and Step II in carbon cycling respectively, and **explain** how each of the processes in Step I and Step II promotes the movement of carbon within the carbon cycle.

Listing the process names and explanation of the process of the movement of carbon in step 1 and step 2: **3pts for step I, 3pts for step II; 6 points maximum**

| Process | Description |
|---|---|
| (1pt) Step I: either photosynthesis or Calvin cycle | (2pts) A chemical process through which plants, some bacteria and algae, produce organic molecules (glucose, $C_6H_{12}O_6$) from CO_2 , using light energy. |
| (1pt) Step II: either (Cellular) Respiration (citric acid cycle/ Krebs cycle) | (2pts) Organic molecules are used/hydrolyzed/broken down, producing CO_2 and ATP in the cells of organisms. |
| (1pt) Step II: fermentation | (2pts) The conversion of organic materials (glucose etc.) to alcohols and CO_2 or organic acids using yeasts, bacteria under anaerobic conditions. |
| (1pt) Step II: decomposition | (2pts) The process of decaying or rotting of dead organic matter, releasing CO_2 . |

| Unidisciplinary question (biology) | Example | Interdisciplinary understanding score |
|---|--|---------------------------------------|
| Fully correct | In step I carbon is taken into plants via photosynthesis. From here it is converted by the plant into organic molecules containing carbon. In step 2 organic molecules are taken up by animals or plants and converted back into carbon dioxide through respiration. | 6 |
| Partially correct | Step I: Through photosynthesis and organic processes, carbon moves into organic molecules such as glucose in plants./Step II: Through transpiration, we remove carbon molecules from our bodies and release it into the atmosphere. | 5 |
| Partially correct | Step I: Carbon is taken out of the atmosphere and locked in organic molecules (2pts) /Step II: Carbon is released from being bound in organic materials as carbon dioxide into the atmosphere (2pts). | 4 |
| Partially correct | In step I plants use CO ₂ to product oxygen through photosynthesis (2pts). In step II humans breathe in oxygen and release CO ₂ to continue the cycle (1pt). | 3 |
| | Step I is carbon being taken up by plants from the soil. Step II is it being eaten and then respired by organisms. | 2 |
| Partially correct | CO ₂ continuously turns into organic molecules, and it is just an everyday thing that is continuous. | 1 |
| Incorrect/Off-topic/Blank/Restatement of the prompt | CO ₂ is absorbed by organic molecules in step 1. Organic molecule breathe out CO ₂ in step 2. | 0 |

Item 4) The paragraph below describes a food chain on the interrelationships that exist between organisms in a field ecosystem. After reading this paragraph, answer the following question.

Oak trees produce large autumnal acorn crops every two to five years. Acorns are a critical food source for the white-tailed deer. Coyotes are important predators of the white-tailed deer. Decomposers consume the remains of dead white-tailed deer, coyotes, and acorns and then return some of nutrients from the dead organic matter back into the soil.

Describe how carbon produced from oak trees is transferred within the food chain through consumers to decomposers and also show how energy is transferred through the food chain.

Explanation of production and movement of carbon in a food chain-

Main point: Carbon moves from plants to animals through food chains. Animals that eat other animals get the carbon from their food too. When plants and animals die, their bodies decay (or are processed by decomposers) bringing the carbon into the air and soil.

4 points maximum

| |
|--|
| (1pt) Through the process of photosynthesis, carbon dioxide is pulled from the air to produce carbon-containing food for acorn growth. |
| (3pts) A white-tailed deer will feed upon the acorn and the carbon (nutrients or CO ₂ -take off 1pt) from the acorn will be transferred to the white-tailed deer (1pt). When a coyote eats the white-tailed deer, carbon moves from the deer to the coyotes. The carbon becomes part of the coyotes (1pt). When the acorns, white-tailed deer, and coyotes die, their remains decay by decomposers, bringing the carbon into the soil and air (1pt). |
| (2pts) In the case of a general description, for example, “carbons are able to move around an ecosystem through the food chain while one organism feeds upon another” |
| (1pt) In the case of a more vague description on carbon movement, for example, “carbons provide for the entire food chain”. |

Description of transfer of energy between trophic levels-

Main point: Energy is transferred through the next trophic levels of a food chain by feeding. In the food chain, the energy that is trapped by oak trees is passed on to the white-tailed deer, and then coyotes.

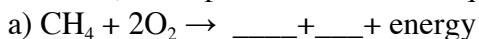
2 points maximum

| |
|--|
| (2pts) Energy is passed from organisms at one trophic level or energy level to organisms at the next trophic level and approximately 10% of the original energy is transferred on to the next trophic level. |
| (1pt) Oak trees' chemical energy is passed on to the white-tailed deer and then the coyote takes in their energy from the white-tailed deer. |
| (1pt) Energy decreases as it moves to higher trophic levels. |

| Disciplinary question (biology) | Example | Interdisciplinary understanding score |
|---------------------------------|--|---------------------------------------|
| Fully correct | Please find the full correct answer! | 6 |
| Partially correct | As the carbon in the acorns falls on the ground, it is then consumed by the white tailed deer. Through this, approximately 10% of the energy is passed onto the deer. Then the coyote eats the deer and 10% of that energy is passed on. When the coyote dies, it is decomposed and the nutrients from its body is then passes through the ground into the soil and through the roots of the oak tree into the tree. | 5 |

| | | |
|-------------------|---|---|
| Partially correct | The Oak tree produces carbon through its acorns. The acorns are eaten by the deer. The deer eat the Coyotes. Decomposers eat the remains of all of the above, returning nutrients to the soil, which can then help boost the ecosystem through more Oak tree production (1.5). Energy follows the same path as carbon up the food chain (0.5) | 4 |
| Partially correct | Carbon, as produced from oak trees are transferred within the food chain, starting with acorns, The acorns from the trees are an important food source for white-tailed deer. Also, as stated, coyotes are important predators of the white-tailed deer, and when dead, their decomposition and acorns can feed on his to continue the cycle. | 3 |
| Partially correct | The carbon produced transferred through the consumption of members of the food chain under the consuming organism. Other compounds that are used as fuel for the production of ATP in organisms are also transferred when consumed by predators. | 2 |
| Partially correct | Energy is transferred through the food chain by organisms eating one another | 1 |
| Partially correct | <ul style="list-style-type: none"> Carbon produced from oak trees could be transferred through nutrients in the soil. As each animal/organism eats other, the carbon and energy goes through the cycle. | 0 |

Item 7) Fossil fuels such as coal, oil and natural gas contain high concentrations of hydrocarbons from the organic remains of prehistoric organisms. The following chemical equation represents the combustion reaction of natural gas (Natural gas is mainly methane). Complete the chemical equation.



Main point: students need to know how to build a basic chemical equation, and products of the complete combustion of Methane and their molecular formula.

4 points maximum

| |
|---|
| $\text{CO}_2 + 2\text{H}_2\text{O}$ (4 pts) |
| $\text{CO}_2 + \# \text{H}_2\text{O}$ if # is not 2 (3 pts) |
| Carbon dioxide and water (2 pts) |
| Either CO_2 or H_2O (no point) |

b) How is the equation similar to or different from the equation for cellular respiration?

Each 2 points for either one similarity or one difference; 4 points maximum

| | Cellular respiration | Combustion |
|--------------|-----------------------------------|------------|
| Similarities | Both produce CO_2 | |
| | Both produce H_2O | |
| | Both need oxygen | |

| | | |
|-------------|---|--|
| Differences | Both involve energy | |
| | Both are chemical changes | |
| | Both are irreversible changes | |
| | It takes place in living cells only. | It does not take place in living cells. |
| | It takes place at the body temperature of the organism. | It takes place at a high temperature than the body temperature of the organism. |
| | Slow process (the oxidation of food and the liberation of energy occur in a step wise manner) | Fast process (The substance is oxidized spontaneously with a sudden release of energy) |
| | A series of interrelated chemical reactions | A single chemical reaction |
| | It is carried out with the help of various enzymes. | Enzymes are not involved in this process. |
| | Energy is liberated in several steps. | Energy is liberated only in one step. |
| | Some energy escapes as body heat and the rest of energy is packaged directly into ATP. | The released energy is dissipated as heat and to some extent as light. |

Note. An answer is vague like “ it has water and CO₂ in it, give 0.5 points.

| Two disciplines question (chemistry and biology) | Example | Interdisciplinary understanding score |
|---|--|---------------------------------------|
| Fully correct | a) CO ₂ + 2H ₂ O b) It is similar in the way it uses a carbon source and oxygen to produce CO ₂ , H ₂ O, and energy. | 8 |
| Partially correct | a) CO ₂ + H ₂ O b) The equation is similar to cellular respiration because both consume oxygen and release carbon dioxide. | 7 |
| Partially correct | a) CO ₂ + 2H ₂ O b) we only release carbon dioxide | 6 |
| Partially correct | a) CO ₂ + 2H ₂ O b) it is similar in that both have CO ₂ outputs | 5 |
| Partially correct | a) CO ₂ + H ₂ + energy b) This equation is similar in that CO ₂ is produced. | 4 |
| Partially correct | a) CO ₂ & H ₂ O / b) cellular respiration is essentially the reverse of this reaction. | 3 |
| Partially correct | a) CH ₄ + H ₂ O b) Energy is also on the product side of the equation for cellular respiration. However, the cellular respiration has CO ₂ on the reactant side where as this chemical equation has O ₂ . | 2 |
| Partially correct | a) H ₂ O + H ₂ b) it is the inverse | 1 |
| Incorrect/Off-topic/Blank/Restatement of the prompt | a) CO ₂ + O ₂ + CH ₄ / b) This equation is different from the natural gas equation because it does not involve methane. a) CH ₄ + O ₂ -> H ₄ + CO ₂ + energy / b) It is different from cellular respiration because it represents the combustion of natural gas. | 0 |

Item 8) Scientists have recently calculated that approximately 26% of all CO₂ emitted from human-related activity, such as the combustion of fossil fuels, was absorbed by oceans during the decade 2002 - 2012. This resulted in 2.5 billion gigatons of excess carbon moved from the atmosphere into the ocean each year over the course of a decade. Scientists are concerned that the mass of CaCO₃ deposited annually in coral reefs is decreasing. They expect that in 2050 the total amount of CaCO₃ in coral reefs will be 20 percent less than it is currently.

Use this information to describe how the combustion of fossil fuels might affect the loss of coral reefs in ocean water, even though the amount of CO₂ in the atmosphere is increasing.

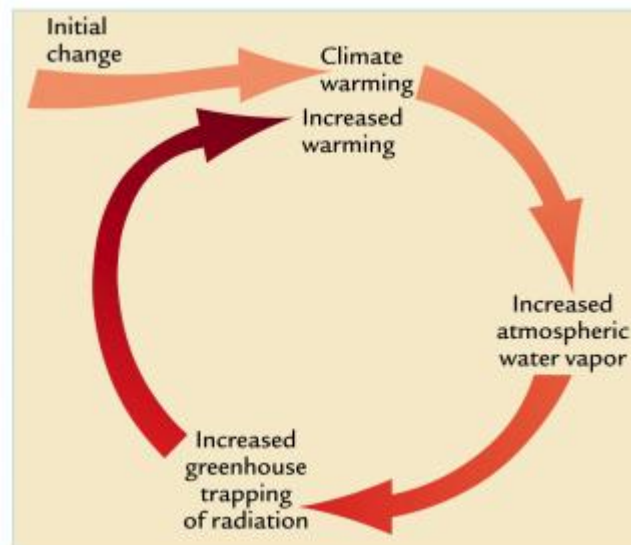
Main point: Increasing CO₂ in the air leads to ocean acidification, which in turn causes slower calcification rates of marine organisms (e.g., reef).

8 points maximum

| |
|---|
| (1pt) Rises in atmospheric carbon dioxide from fossil fuel combustion increase the concentration of carbon dioxide in the air. |
| (1pt) The increasing amount of CO ₂ in the atmosphere moves into the oceans (by Henry's law that states that the solubility of a gas in a liquid is directly proportional to the partial pressure of the gas over the liquid). |
| (2pts) In the oceans, CO ₂ reacts to generate carbonic acid (H ₂ CO ₃), which consequently releases hydrogen ions to form bicarbonate (HCO ₃ ⁻) and carbonate ions (CO ₃ ²⁻). |
| (2pts) The increased concentration of hydrogen ions lowers the pH in the oceans. |
| (2pts) Lower pH affects the equilibrium between bicarbonate (HCO ₃ ⁻) and carbonate ions (CO ₃ ²⁻), decreasing proportion of carbonate (or CaCO ₃). |
| (2pts) The combustion of fossil fuels (or the increasing amount of CO ₂ in the atmosphere) leads to ocean acidification. |
| (2pts) The combustion of fossil fuels (or the increasing amount of CO ₂ in the atmosphere) leads to a decrease in CaCO ₃ and thus, a decrease in coral reef production. |
| (2pts) The increase in acidity decreases the total amount of coral reefs. |
| (1pt) Carbonate ions (or CaCO ₃) are needed for certain coral animals to make the coral's structures. |
| (2pts) The reduction in the concentration of carbonate ions leads to an increase in reef dissolution rates, which in turn leads to loss of coral reef or the reduction in the concentration of carbonate ions results in a decrease in rates of calcification (production of reef structure), which eventually leads to loss of coral reef. |
| (1pt) The combustion of fossil fuels (or the increasing amount of CO ₂ in the atmosphere) results in a decrease of the amount of coral reefs (or CaCO ₃). |

| Two disciplines question (chemistry and environmental science) | Example | Interdisciplinary understanding score |
|--|---|---------------------------------------|
| Fully correct | Combustion of fossil fuels creates an excess of CO ₂ in the atmosphere. The oceans are carbon sink for this excess CO ₂ . The increase of CO ₂ in the water creates a weak acid in the oceans. As the acid is created, calcium carbonate is depleted. With the depletion of calcium carbonate, coral reefs will suffer. | 8 |
| Partially correct | The combustion of fossil fuels increases the amount of carbon released into the atmosphere. This will lead to increase in mixing of it in the ocean. This will increase the acidity of the ocean water. This increase in acidity could kill the fish this resulting in the loss of the coral reefs. | 6 |
| Partially correct | The combustion of fossil fuels produces a significant amount of CO ₂ . The amount of CO ₂ produced is too much for the ocean to absorb, and there are not enough bicarbonate ions to compensate for the amount of CO ₂ being absorbed in the ocean. This causes the acidity of the ocean to change which can affect the coral. | 5 |
| Partially correct | The combustion of fossil fuels releases CO ₂ into the atmosphere. Ocean water is a sequestration environment for CO ₂ from the atmosphere, the more CO ₂ that gets sequestered the fewer the amount of CaCO ₃ that gets deposited in coral reefs each year. | 4 |
| Partially correct | Coral is an animal. When CO ₂ is released to the atmosphere, some of it is absorbed by the water. Because there is CO ₂ dissolved in the water now, the coral reefs are starting to “suffocate” basically and die meaning there is less coral to make CaCO ₃ . | 3 |
| Partially correct | CO ₂ might be increasing but the CaCO ₃ is decreasing thus killing off coral reefs. | 2 |
| Partially correct | With more and more combustion of fossil fuels, there will be less and less coral reefs. | 1 |
| Incorrect/Off-topic/Blank/Restatement of the prompt | <ul style="list-style-type: none"> The burning of fossil fuels could release other noxious gases that could be harmful to the reefs. Increase global warming, increases the standard ocean temp, killing coral reefs. | 0 |

Item 12) The figure below shows positive feedback cycle. In the context of global warming,



a) What could the initial change be in this process? **4 points maximum**

| |
|---|
| (4pts) Increased atmospheric CO ₂ |
| (4pts) Anthropogenic sources (e.g., burning fossil fuels, burning forests, deforestation and destruction of the soil etc.) |
| (4pts) Increased other greenhouse gas emissions (e.g., methane (CH ₄), water vapor (H ₂ O), nitrous oxide (N ₂ O), Fluorinated gases) in the atmosphere |
| (2pts) Change in CO ₂ levels or amount of CO ₂ in the atmosphere |
| (2pts) Human activities |

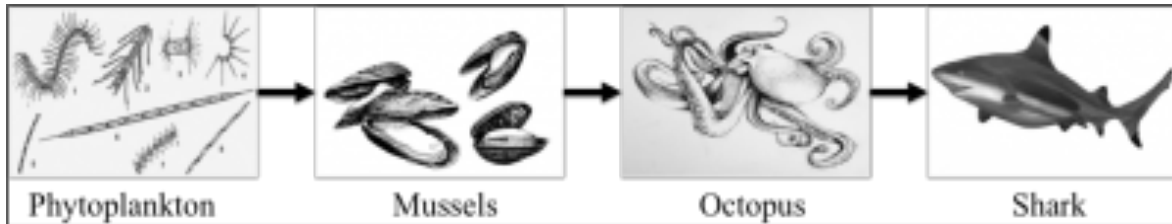
b) Discuss how the initial change leads to the feedback loop in the diagram below.
Main point: As the Earth's temperature rises from a given initial forcing, the positive feedback in the system results in additional warming. **1 point each row; 4 points maximum**

| |
|---|
| 1. One of initial changes mentioned above causes climate warming. |
| 2. which in turn increases the air's capacity to hold more water vapor. |
| 3. Water vapor is a potent greenhouse gas |
| 4. And it helps the Earth hold on to more infrared (IR) radiation reflected back to the Earth by the atmosphere, which in turn warms the climate further. |
| It is non-stop (continuous) cycle. |

Note. If the order was changed, no points.

| Q12) Two disciplines question (physics and earth science) | Example | Interdisciplinary understanding score |
|---|---|---------------------------------------|
| Fully correct | a) Increase of CO ₂ in the atmosphere/ b) with an increase of CO ₂ , more global warming occurs. This leads to more evaporation and an increase of water vapor in the atmosphere. Water vapor is considered a greenhouse gas and radiates back the heat it traps in. This in turn causes more warming. | 8 |
| Partially correct | a) increase CO ₂ in atmosphere b) increased CO ₂ will affect the overall climate of the earth by increasing the temperature. Once the temperature is higher, more water will be absorbed as well, along with other greenhouse gases that will further change the climate. | 7 |
| Partially correct | The initial change is a huge release of CO ₂ in the form of humans burning fossil fuels. That influx in CO ₂ causes the atmosphere to trap more heat, which warms the planet more and creates more greenhouse gases. | 6 |
| Partially correct | a) increase in atmospheric CO ₂ levels b) when there is an initial increase in atmospheric CO ₂ it allows for global warming and allows the other processes to increase as well. | 5 |
| Partially correct | a) increase in atmospheric CO ₂ | 4 |
| Partially correct | Increasing temperatures will lead to greater amounts of evaporation (higher temp = greater/faster chemical reactions), which leads greater building up of greenhouse gases in the atmosphere, which leads to more trapped heat on Earth due to the greenhouse effect, which thus leads to increased temperatures, sealing a positive feedback loop. | 3 |
| Partially correct | Increased levels of CO ₂ in the atmosphere. / It traps the heat from the sun in the atmosphere making those things happen. | 2 |
| Partially correct | a) the initial change could be the increase of pollution / b) the increase of pollution which leads to increased greenhouse gases and then increased temperature. | 1 |
| Incorrect/Off-topic/Blank/Restatement of the prompt | a) Increased sun radiation / b) Well if it happens, it'll affect everything else in the diagram | 0 |

Item 14) A Figure below shows one example of a food chain in an ocean ecosystem. The emission of anthropogenic CO₂ into the atmosphere causes changes to ocean ecosystems. How could the increased amount of CO₂ in the atmosphere affect the food chain of the ocean ecosystem?



Main point: Links between direct effects of ocean acidification at the organism level and indirect effects on food web structure and ecosystem functioning need to be explained. A disturbance of marine ecosystem such as ocean acidification leads to cascading effects throughout the marine ecosystem, leading to changes in predator-prey interactions.

8 points maximum

(2 pts) Discussing the relationship between atmospheric CO₂ and oceans; When the oceans absorb CO₂, the chemical reaction that takes place produces carbonic acid (H₂CO₃), which increases the acidity (lowers the pH) of seawater (chemistry).

| |
|--|
| (3pts) Discussing the effect of lower pH on phytoplankton; The lower pH in the oceans may lead to metabolism changes (e.g., respiration, photosynthesis, and nutrient dynamics) of phytoplankton, and consequently, limits for growth (or cell volume) (environmental science). |
| (3pts) Discussing the effect of lower pH on phytoplankton; Ocean acidification can enhance the growth of plankton population because of increasing photosynthetic activity (environmental science). |
| (2pts) Discussing the effect of increased CO ₂ level on phytoplankton: increased CO ₂ leads to less phytoplankton. |
| (3pts) Discussing the effect of lower pH on mussels. Example: The acidic environment interferes with the ability to make mussels' calcium carbonate shells increase dissolution of the carbonated shells (chemistry and environmental science). |
| (2pts) Discussing the effect of increase CO ₂ level on mussels; Increased CO ₂ leads to fewer mussels. |
| (3pts) Discussing the impact of change in the size of phytoplankton population on the natural competition related to higher tropic level. Example: It could cause more phytoplankton to be produced, increasing the number of all of the predators in the food chain above (biology). |
| (3pts) Discussing the impact of change in the size of mussel population on the phytoplankton, the octopus, and the shark. Example: These changes among small creatures at the bottom of the food chain could have significant effects on the quantity or composition of octopus and shark at the top (biology). |
| (2pts) Mentioning more broad words on the disruption of food chain in marine ecosystem. Example 1: These changes among small creatures at the bottom of the food chain could have significant effects on ecosystem structure, biodiversity, and ultimately ecosystem health. Example 2: If the ocean becomes acidic, then marine life would be detrimentally impacted. |
| (1pt) If the response' reasoning is vague, for example, "Increased CO ₂ could lead to some fish getting killed off", or "Increased CO ₂ causes water acidity so animals cannot survive as well". |

| Interdisciplinary question (biology, chemistry and environmental science) | Example | Interdisciplinary understanding score |
|---|---|---------------------------------------|
| Fully correct | Increased amount of CO ₂ in the atmosphere means that the amount of CO ₂ in the oceans also increase. Increased CO ₂ caused the increased ions of H ⁺ and HCO ₃ in the oceans which make the water more acidic. This contributes to the deaths of mussels, whose shells are compromised in acidic water. This would increase the population of phytoplankton who are preyed on by mussels. It would also decrease the population of octopus unless they found other prey. If the octopus population did decrease, the shark population may also decrease unless they found other prey. | 8 |
| Partially correct | An increased amount of atmospheric CO ₂ could dissolve into the water and acidify it, making conditions less than optimal for the phytoplankton which would cause them to leave or die off therefore reducing the population of mussels, octopus, and shark. | 7 |
| Partially correct | An increase in carbon dioxide will cause an increase in phytoplankton. Since these will be in excess it could lead to an increase in the population of all of these animals. It could also lead to an increase in completion and add a new organism because of the increase in food availability. | 5 |
| Partially correct | The increased amount of CO ₂ affects the amount of CO ₂ available to phytoplankton, who use the compound as a fuel in photosynthetic respiration. Because the phytoplankton are affected, the rest of the food chain is affected. | 4 |
| Partially correct | Phytoplankton populations would increase leading to a surplus of food. This would cause a greater number of each organism in the chain. | 3 |
| Partially correct | This would impact the entire ecosystem: as the smallest food source would be impacted by increased levels of CO ₂ & be carried through the food chain as it is consumed. | 2 |
| Partially correct | One of the main food groups could die off causing the other groups to suffer and eventually die off as well. | 1 |
| Incorrect/Off-topic/Blank/Restatement of the prompt | <ul style="list-style-type: none"> • The pressure of CO₂ in the water would also increase. This would most likely hurt the beginning of the food chain in turn affecting everything that comes after. • Because CO₂ is very important at the biosphere level. | 0 |

Item15) The geological carbon cycle is complicated, with many different pieces playing their roles. Usually a change in one part of the cycle causes compensating changes in other parts, but sometimes the system takes a long time to get back into balance. For example, it takes a long time for the oceans to increase their uptake of carbon dioxide, so

they might not be able to compensate for CO₂ increase in the air, resulting in an imbalance. How could deforestation lead to an imbalance in carbon dioxide levels?

Main point: The role of Forest in the carbon cycling and effects of deforestation on the carbon cycle should be discussed.

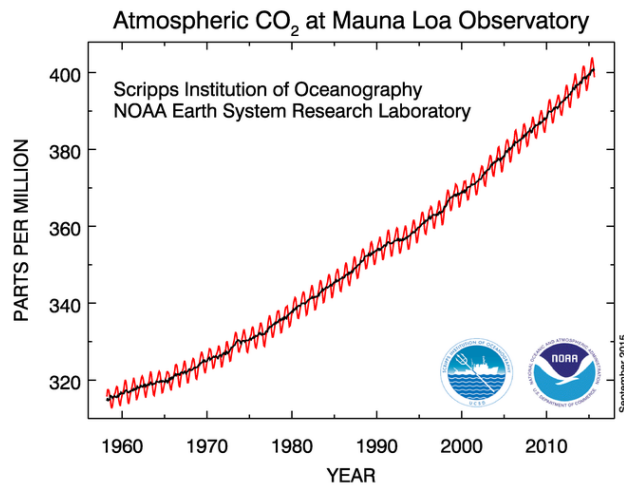
8 points maximum

| |
|--|
| (4pts) "Slash and burn" technique used to clear the forest releases large amounts of carbon dioxide (or carbon) into the atmosphere when biomass burns. |
| (4pts) Deforestation reduces the amount of photosynthesis (2pts); thus, it increases CO ₂ level in the atmosphere (2pts). |
| (4pts) Deforestation accelerates the decomposition rate (2pts), thus, it increases CO ₂ level in the atmosphere (2pts). |
| (2pts) Trees remove CO ₂ from the atmosphere. Plants absorb (consume) CO ₂ without stating the CO ₂ level. |
| (2pts) Deforestation increases CO ₂ level in the atmosphere without reasoning. |
| (2pts) Mentioning the role of forest and plants and the increase level of CO ₂ ; for example, forests are carbon sink, trees remove CO ₂ from the atmosphere, or plants absorb (consume) CO ₂ . |
| (3pts) Less intake of CO ₂ leads to an increase of CO ₂ level in the atmosphere. |

Note. If a student used just "imbalance of CO₂ level without expression of increased level of CO₂, take off 1pt.

| Interdisciplinary question (biology, chemistry) | Examples | Interdisciplinary understanding score |
|---|--|---------------------------------------|
| Fully correct | Trees take in CO ₂ during photosynthesis. The removal of forests will lead to an increase in atmospheric CO ₂ concentrations. The larger the plant, the more CO ₂ it takes in (environmental science). Also practices such as slashing and burning can lead to more CO ₂ being produced from the fire. | 8 |
| Partially correct | Deforestation causes more CO ₂ to be in the atmosphere, as the levels of CO ₂ increase in the atmosphere, there is an imbalance in the system, as there isn't enough plants to compensate for the increased level of CO ₂ . Because of this, a balance in the system may take a long time to occur. | 6 |
| Partially correct | Deforestation would increase the amount of carbon in the atmosphere since there will be no trees to absorb the carbon. | 4 |
| Partially correct | Large release of CO ₂ into atmosphere means that the ocean would have to catch up on sequestering the CO ₂ . The atmosphere would have more CO ₂ than the ocean could originally sequester. | 3 |
| Partially correct | Deforestation would release more CO ₂ in the air, but also causes an imbalance because O ₂ would not be released as much as it would have with the forest. | 2 |
| Partially correct | Because no trees means less plants to take in the oxygen | 0 |

Item 19) The graph below shows measurements of carbon dioxide in the atmosphere (measured in parts per million) taken at a NOAA station on Hawaii between 1958 and 2013. The red line represents a shorter-time scale variation of the monthly mean values of CO₂ in a given year. The black line represents a long-term trend over the last decades, after correction for the average seasonal cycle.



Some people claim that the fluctuations in the atmospheric CO₂ level are due to variations in radiation coming from the Sun.

a) Describe how the amount of incident sunlight might affect the amount of carbon dioxide in the atmosphere.

Main point: Relationship between the change of amount of sunlight in seasonal change and photosynthetic activity of plants allows fluctuating carbon dioxide (CO₂) level in the atmosphere.

4 points maximum

| |
|--|
| (2pts) The more sunlight, the more CO ₂ in the atmosphere, and vice versa. |
| (2pts) Amount of sunlight affects photosynthesis rates. |
| (4pts) More sunlight allows plants to take up more of CO ₂ during photosynthesis activity, decreasing the CO ₂ levels in the atmosphere. |
| (4pts) Less sunlight reduces the rate of intake of CO ₂ from the atmosphere. |

b) Are both of the red line and the black line pattern equally likely to be due to sunlight variation? Explain why or why not.

Main point: the reasons on different pattern shown in the red and the black line.

4 points maximum

| |
|--|
| (4pts) No, the red line's fluctuating pattern is due to seasonal solar variation. The increasing pattern of the black line over time is due to human activities that increase CO ₂ in the air (e.g., fossil fuel emissions or deforestation etc.). If the reasoning about photosynthesis is vague, for example, the more sunlight, the more plants grow.....,take off 1pt. |
|--|

| |
|---|
| (3pts) No, the red line's fluctuating pattern is due to seasonal solar variation. The black line looks like a trend caused by something else. |
| (2pts) No, the red line's fluctuating pattern is due to seasonal solar variation. |
| (2pts) No, the black line is affected by human activities that increase CO ₂ in the air. |
| (1pt) No, the red line is on a short-term scale and whereas the black line is on a long-term scale. |

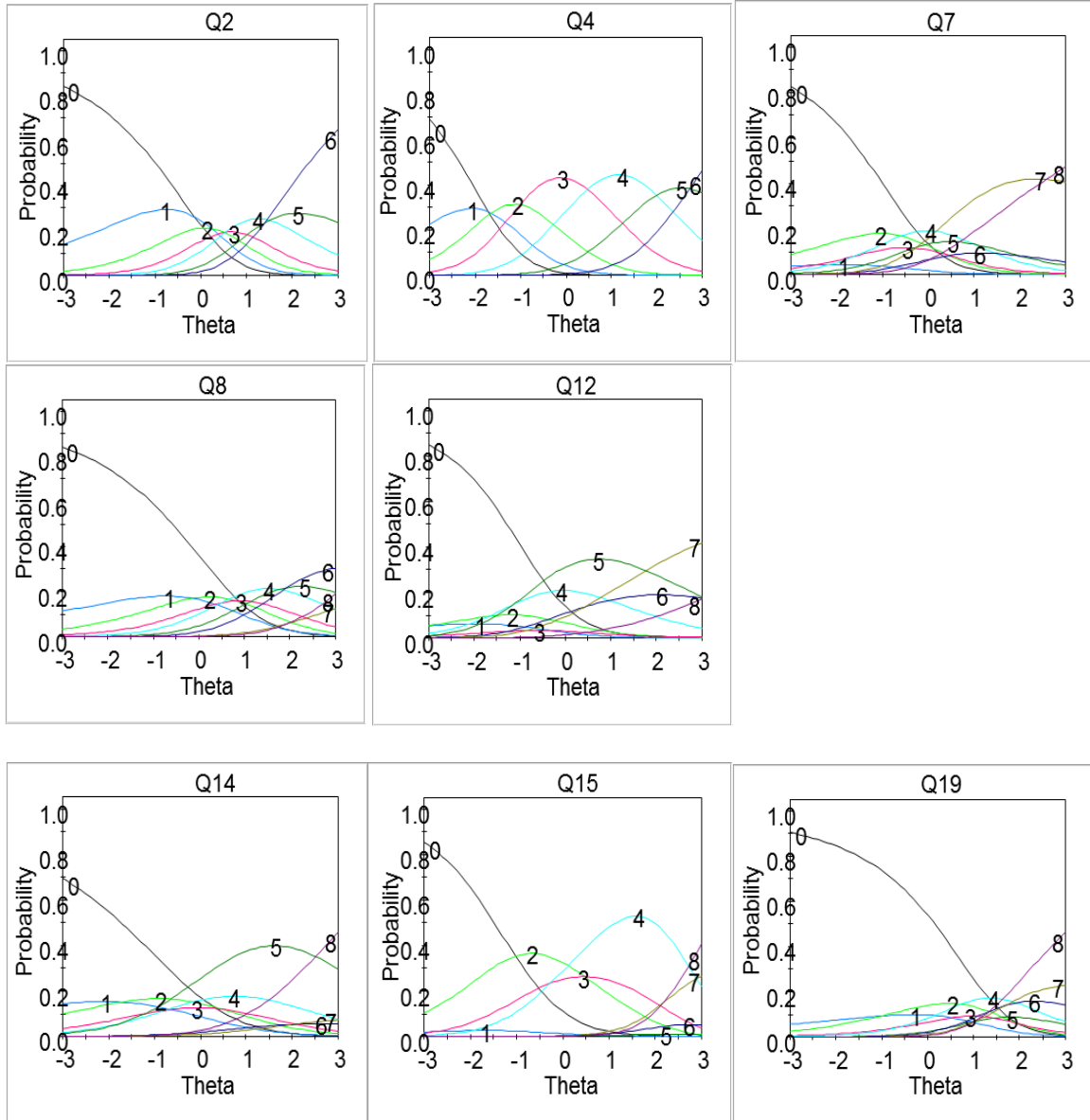
Note. Even though one answered “no” in (b), if the reason is not correct or relevant, no points. For black lines question, if only expressed “other factor”, not mentioning a concrete example, 1pt.

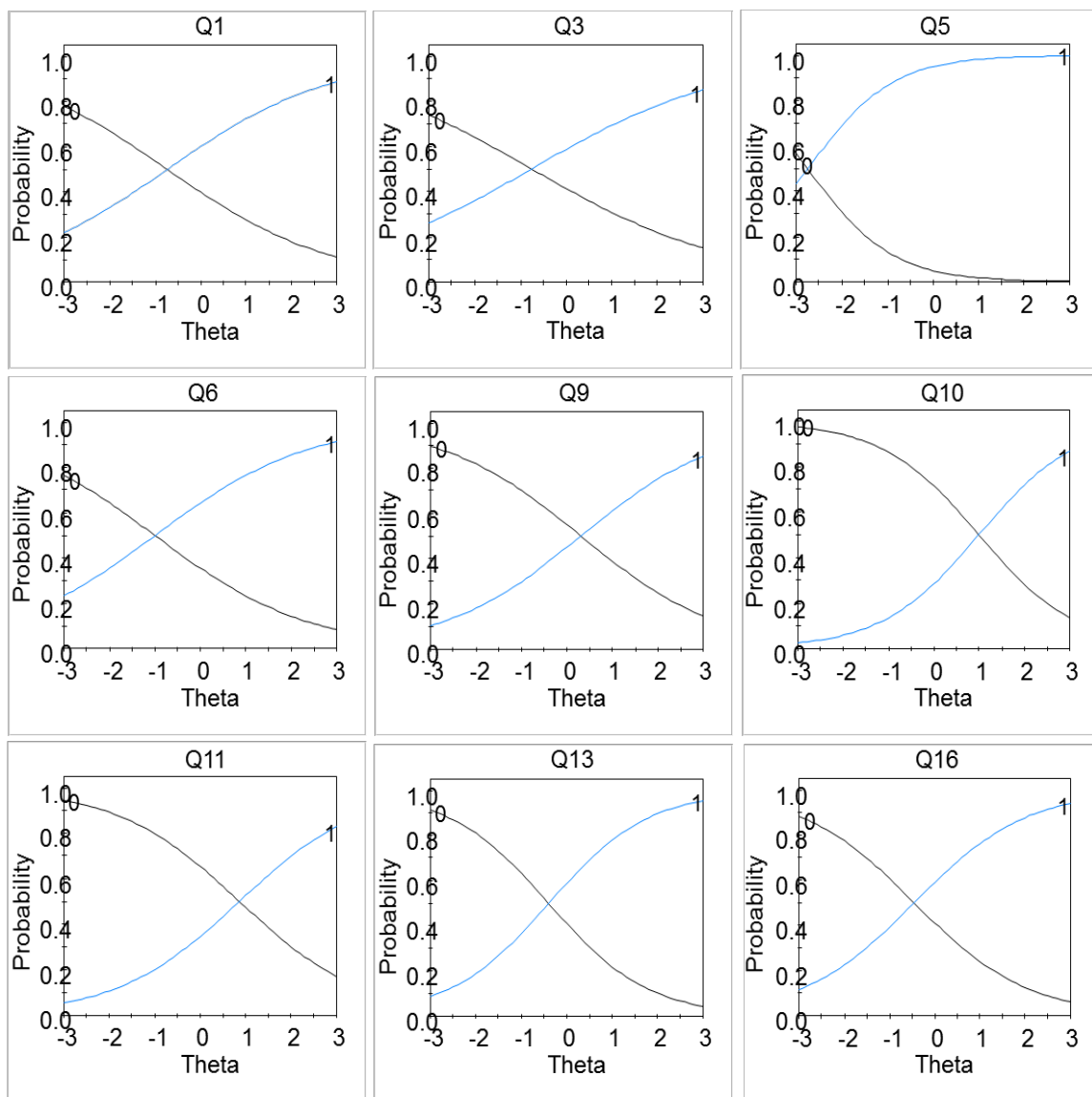
| Q19) Interdisciplinary (biology and earth science) | Example | Interdisciplinary understanding score |
|--|--|---------------------------------------|
| Fully correct | The amount of incident light controls how much light energy plants are getting to carry out photosynthesis. An increase in light energy leads to an increase of photosynthetic processes and a decrease in the CO ₂ levels in the atmosphere. b) No because these solar variations are short term, so it would only show up in the red line. The black line is more for an increase in deforestation and fossil fuel burning over time. | 8 |
| Partially correct | a) the sunlight provide the energy for plants to convert CO ₂ in the atmosphere into the sugar. So more sunlight would result in the depletion of CO ₂ in the atmosphere. b) the red line are more likely to be due to solar variations since it is more seasonally cyclic. On the other hand, black line shows the increasing CO ₂ which is likely to be the result from other factor. | 7 |
| | a) When there is more incident sunlight, plants are capable of more photosynthesis, thereby decreasing the amount of CO ₂ in the atmosphere. b) The red line is likely to fluctuation due to seasonal variation in incident sunlight due to earth's tilt, but the gradual increase of the black and red lines are not likely caused by the sun, because there would be no process that would gradually be increasing the amount of sunlight we receive. | 5 |
| Partially correct | a) less sun= rise in CO ₂ . This is because a decrease in photosynthesis which consumes CO ₂ . B) just the black line. An overall change in solar variation needs to be recorded over a long period of time. | 4 |
| Partially correct | a) The amount of incident sunlight coming to Earth affects the amount of carbon dioxide in the atmosphere, because the amount of sunlight affects how well plants can convert CO ₂ to O ₂ . b) No, because the red line represents shorter amounts of time,, while sunlight should affect the CO ₂ in a longer term time period, the blank line. | 3 |

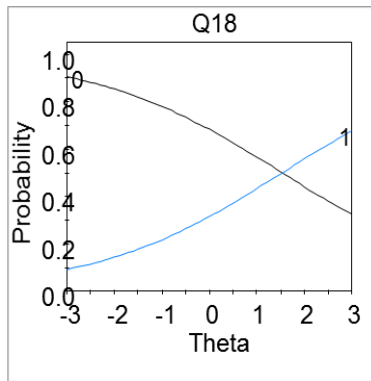
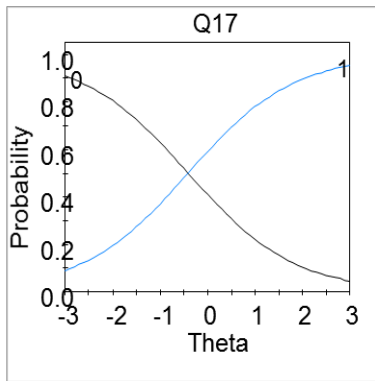
| | | |
|---|---|---|
| Partially correct | a) the amount of sunlight affects the ability of plant and other organisms to grow, therefore affecting amount of CO ₂ in atmosphere. | 2 |
| Partially correct | a) increased sunlight =increased global temperatures on Earth =feedback loops cause increase in CO ₂ . b) Most likely solar variations will not cause such a steep increase in CO ₂ . | 1 |
| Incorrect/Off-topic/Blank/Restatement of the prompt | The amount of incident sunlight may affect the amount of carbon dioxide in the atmosphere as it only could add to the warming of the planet, and it adds to the trapped greenhouse gases in our atmosphere (CO ₂). The red and black lines may be equally likely as it could take into account the amount of incident sun and parts per million of CO ₂ in the atmosphere. | 0 |

Appendix C: Item characteristic and item information curves

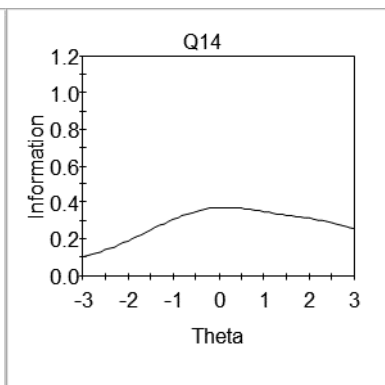
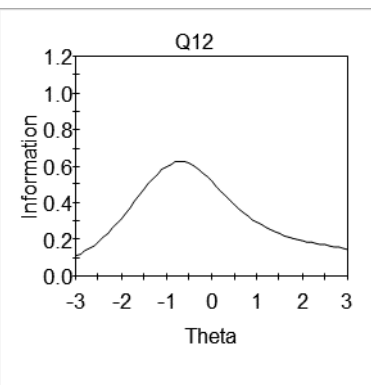
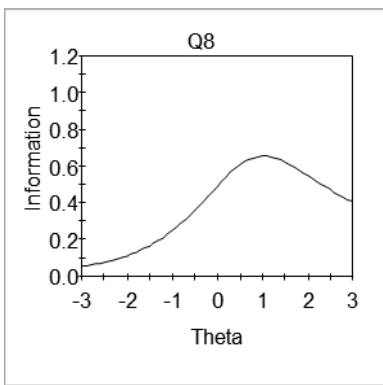
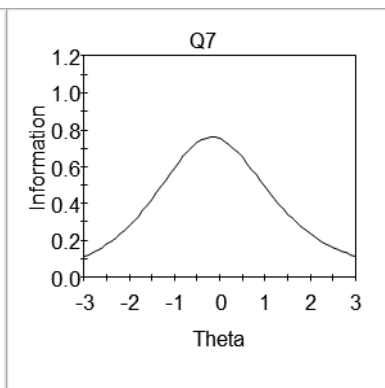
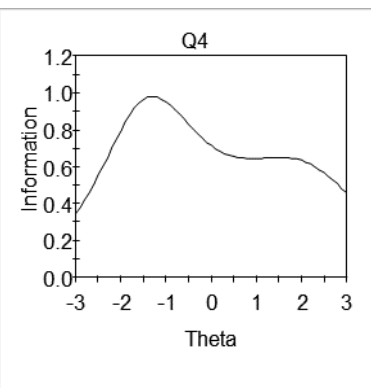
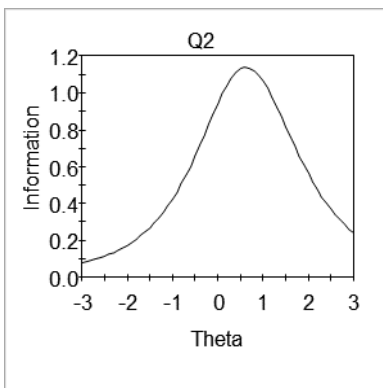
Item characteristic curve

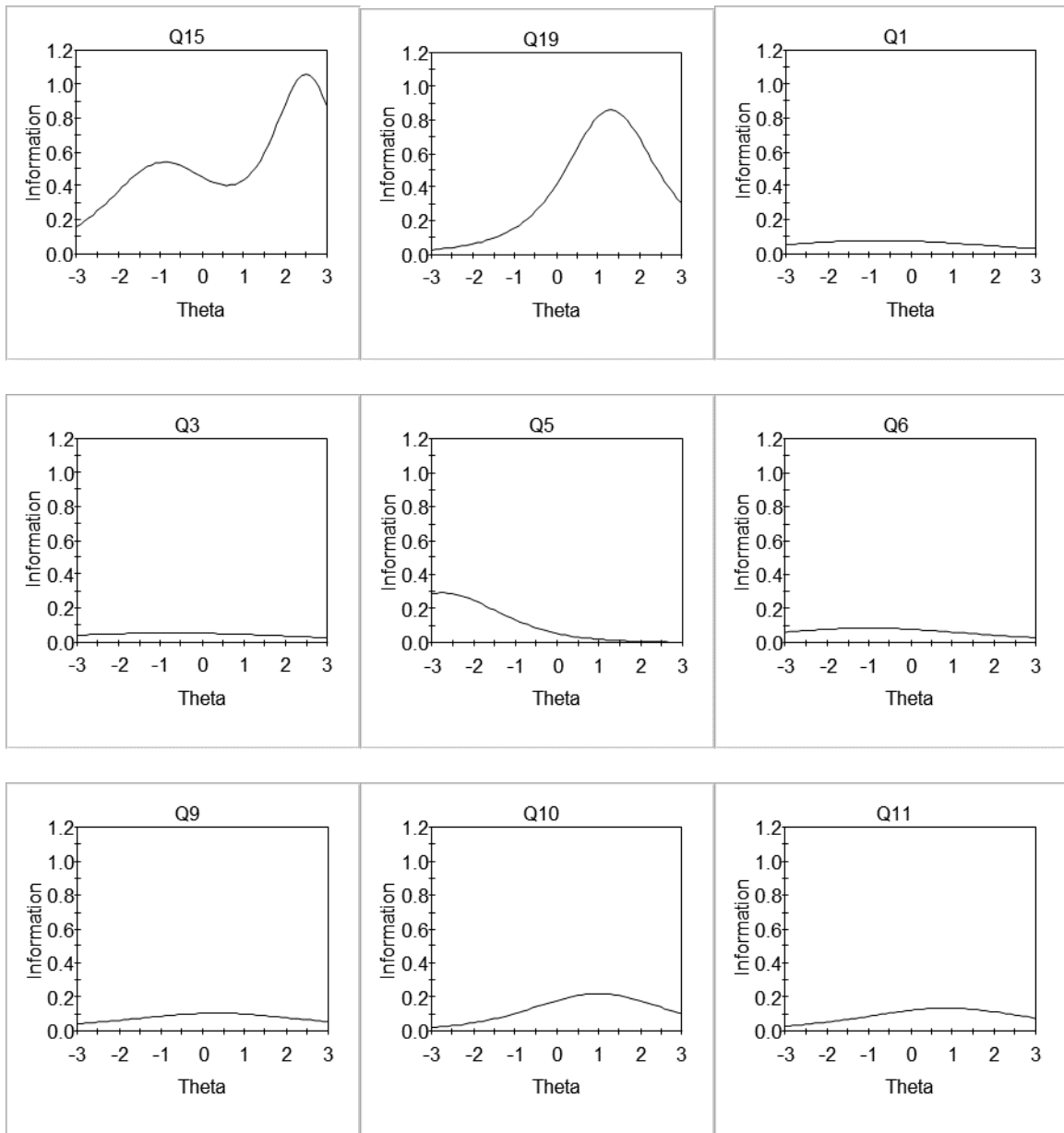


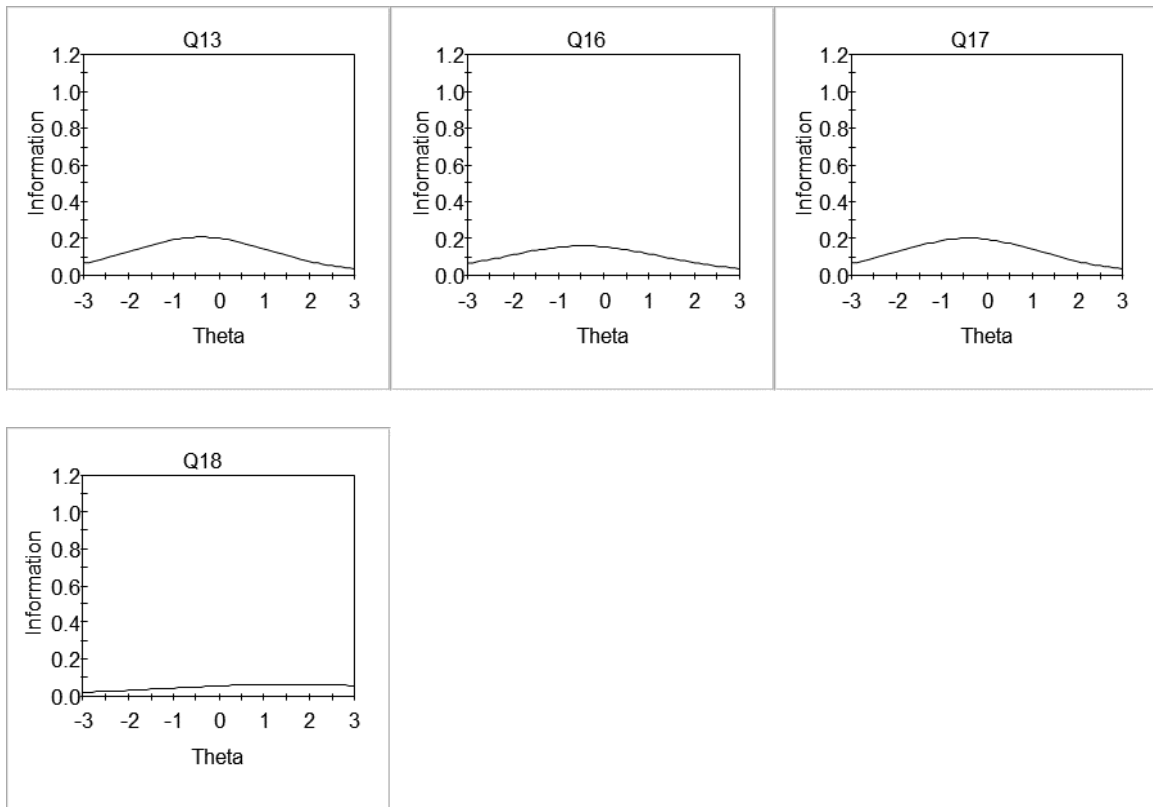




Item information curve







References

- Aggarwal, A. (2003). *Web-based education: Learning from experience*. Hershey, PA: IGI Global.
- American Association for the Advancement of Science (AAAS). (1989). *Science for all Americans*. Washington, DC: American Association for the Advancement of Science.
- American Association for the Advancement of Science (AAAS). (2009). *Benchmarks for science literacy on-line*. Retrieved from <http://www.project2061.org/publications/bsl/online>.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Astin, A. W. (1993). *What matters in college?: Four critical years revisited*. Jossey-Bass San Francisco.
- Baker, F. B. (2001). *The basics of item response theory* (2 ed.). ERIC Clearinghouse on Assessment and Evaluation.
- Balsiger, P. W. (2004). Supradisciplinary research practices: History, objectives and rationale. *Futures*, 36(4), 407-421.
- Barab, S. A., & Landa, A. (1997). Designing effective interdisciplinary anchors. *Educational Leadership*, 54(6), 52-55.
- Baxter, G., & Mislevy, R. J. (2004). *The case for an integrated design framework for assessing science inquiry*. Retrieved from
- Beane, J. A. (1995). Curriculum integration and the disciplines of knowledge. *Phi Delta Kappan*, 616-622.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores*. MA: Addison-Wesley.
- Boix Mansilla, V. (2005). Assessing student work at disciplinary crossroads. *Change: The Magazine of Higher Learning*, 37(1), 14-21.
- Boix Mansilla, V. (2006). Assessing expert interdisciplinary work at the frontier: An empirical exploration. *Research Evaluation*, 15(1), 17-29.
- Boix Mansilla, V., & Duraisingh, E. D. (2007). Targeted assessment of students' interdisciplinary work: An empirically grounded framework proposed. *The Journal of Higher Education*, 78(2), 215-237.
- Boix Mansilla, V., Miller, W. C., & Gardner, H. (2000). On disciplinary lenses and interdisciplinary work. In S. Wineburg & P. Gossman (Eds.), *Interdisciplinary*

- curriculum: Challenges to implementation* (pp. 17-38). New York: Teachers college, Columbia University.
- Bond, L. A. (1995). *Critical issue: Rethinking assessment and its role in supporting educational reform*. Oaks Brooks, IL: North Central Regional Education Laboratory.
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer Netherlands.
- Bragaw, D., Bragaw, K. A., & Smith, E. (1995). Back to the Future: Toward curriculum integration. *Middle School Journal*, 27(2), 39-46.
- Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn : Brain, mind, experience, and school*. Washington, DC: National Academy Press.
- Braunger, J., & Hart-Landsberg, S. (1994). *Crossing boundaries: Explorations in integrative curriculum*. Portland, OR: Northwest Regional Educational Laboratory.
- Bresciani, M. J., Zelna, C. L., & Anderson, J. A. (2004). *Assessing student learning and development*. United States: NASPA.
- Buchbinder, S. B., Alt, P. M., Eskow, K., Forbes, W., Hester, E., Struck, M., & Taylor, D. (2005). Creating learning prisms with an interdisciplinary case study workshop. *Innovative Higher Education*, 29(4), 257-274.
- California Department of Education. (1990). *The California framework for science instruction*. Sacramento, CA: California Department of Education.
- Chandramohan, B., & Fallows, S. J. (2009). *Interdisciplinary learning and teaching in higher education: Theory and practice*. New York, NY: Routledge.
- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265-289.
- Chi, M. T. H., & Bassok, M. (1989). Learning from examples via self-explanations. *Knowing, Learning, and Instruction: Essays in honor of Robert Glaser*, 251-282.
- Chi, M. T. H., & Ceci, S. J. (1987). Content knowledge: Its role, representation, and restructuring in memory development. In H. W. Reese (Ed.), *Advances in child development and behavior* (Vol. 20, pp. 91-142). New York: Academic Press.
- Chi, M. T. H., E, H. J., & Robin, A. F. (1988). *Knowledge-constrained inferences about new domain-related concepts: Contrasting expert and novice children*. Pittsburgh University, PA: Learning Research and Development Center.
- Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ: L. Erlbaum Associates.
- Choi, B., & Pak, A. (2006). Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. *Clinical and investigative medicine*, 29(6), 351-364.

- Chynoweth, P. (2009). The built environment interdiscipline: A theoretical model for decision makers in research and teaching. *Structural Survey*, 27(4), 301-310.
- Clarke, J. H., & Agne, R. M. (1997). *Interdisciplinary high school teaching: Strategies for integrated learning*. Boston: Allyn & Bacon.
- Clary, R. M., & Wandersee, J. H. (2007). A mixed methods analysis of the effects of an integrative geobiological study of petrified wood in introductory college geology classrooms. *Journal of Research in Science Teaching*, 44(8), 1011-1035.
- College Board. (2009). *Science: College board standards for college success*. New York: College Board
- Collins, H., & Evans, R. (2007). *Rethinking expertise*: University of Chicago Press.
- Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of consulting and clinical psychology*, 56(5), 754.
- Council, N. R. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas* (J. W. Pellegrino, M. R. Wilson, J. A. Koenig, & A. S. Beatty Eds.). Washington, D.C: The National Academies Press.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: CBS College Publishing.
- Davis, J. R. (1995). *Interdisciplinary courses and team teaching: New arrangements for learning*: American Council on Education and the Oryx Press Phoneix, AZ.
- de Baar, H. J., & Suess, E. (1993). Ocean carbon cycle and climate change—An introduction to the interdisciplinary union symposium. *Global and Planetary Change*, 8(1-2), VII-XI.
- Dewey, J. (1938). *Experience and education*. New York: The Macmillan company.
- Doll, W. E. (1993). *A post-modern perspective on curriculum*. New York: Teachers College Press.
- Drake, S. M., & Burns, R. C. (2004). *Meeting standards through integrated curriculum*. Alexandria, Va: Association for Supervision and Curriculum Development.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New York: Psychology Press.
- Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (2006). *The Cambridge handbook of expertise and expert performance*: Cambridge University Press.
- Ericsson, K. A., Nandagopal, K., & Roring, R. W. (2009). An expert performance approach to the study of giftedness. In L. Shavinina (Ed.), *International handbook on giftedness* (pp. 129-153). Berlin: Springer.
- Field, M., Lee, R., & Field, M. L. (1994). Assessing interdisciplinary learning. *New Directions for Teaching and Learning*, 1994(58), 69-84.
- Fleiss, J. L. (1986). Reliability of measurement. *The design and analysis of clinical experiments*, 1-32.

- Foss, D. H., & Pinchback, C. L. (1998). An interdisciplinary approach to science, mathematics, and reading: Learning as children learn. *School Science and Mathematics*, 98(3), 149-155.
- Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: an introduction*. Thousand Oaks, CA: Sage.
- Ganaras, K., Dumon, A., & Larcher, C. (2008). Conceptual integration of chemical equilibrium by prospective physical sciences teachers. *Chemistry Education Research and Practice*, 9(3), 240-249.
- Gehrke, N. J. (1998). A look at curriculum integration from the bridge. *Curriculum Journal*, 9(2), 247-260.
- Golding, C. (2009). *Integrating the disciplines: Successful interdisciplinary subjects*. Melbourn: Centre for the Study of Higher Education, University of Melbourne.
- Haladyna, T. M. (2006). Perils of standardized achievement testing. *Educational Horizons*, 85(1), 30-43.
- Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability correlation versus percentage of agreement. *Written Communication*, 16(3), 354-367.
- Henson, K. T. (2003). Foundations for learner-centered education: A knowledge base. *Education*, 124(1), 5-16.
- Hinkin, T. R., Tracey, J. B., & Enz, C. A. (1997). Scale construction: Developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research*, 21(1), 100-120.
- Hirsch, E. D. (1996). *The schools we need and why we don't have them*. New York: Doubleday.
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6(1), 53-60.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1), 1-55.
- Hurd, P. D. (1991). Why we must transform science education. *Educational Leadership*, 49(2), 33-35.
- Hursh, B., Haas, P., & Moore, M. (1983). An interdisciplinary model to implement general education. *The Journal of Higher Education*, 42-59.
- Ivanitskaya, L., Clark, D., Montgomery, G., & Primeau, R. (2002). Interdisciplinary learning: Process and outcomes. *Innovative Higher Education*, 27(2), 95-111.
- Jacobs, H. H. (1989). *Interdisciplinary curriculum: Design and implementation*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Jantsch, E. (1947). Inter - and transdisciplinary university: A systems approach to education and innovation. *Higher Education Quarterly*, 1(1), 7-37.

- Jantsch, E. (1971). Inter-and transdisciplinary university: A systems approach to education and innovation. *Policy Sciences*, 1(1), 430-437.
- Johnston, J., Riordain, M. N., & Walshe, G. (2014). An integrated approach to the teaching and learning of science and mathematics utilizing technology-The teachers' perspective. *Journal on School Educational Technology*, 9(4), 14-26.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4 ed., pp. 17-64). New York: American Council on Education/Macmillan.
- Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching*, 50(5), 530-560.
- Klein, J. T. (1990). *Interdisciplinarity: History, theory, and practice*. Detroit, MI: Wayne State University Press.
- Klein, J. T. (2002). *Interdisciplinary education in K-12 and college: A foundation for K-16 dialogue*. New York, NY: College Board Publications.
- Klein, J. T., & Newell, W. (1997). Advancing interdisciplinary studies. In J. Gaff & J. Ratcliff (Eds.), *Handbook on the Undergraduate Curriculum*. (pp. 393-415). San Francisco: Jossey-Bass.
- Kliebard, H. M. (2004). *The struggle for the American curriculum, 1893-1958*. New York, NY: Routledge.
- Kline, R. (2015). *Principles and Practice of Structural Equation Modeling* (4th ed.). New York: Guilford Publications.
- Knapp, E., Desjardins, S., & Pleva, M. (2003). An interdisciplinary approach to teaching introductory chemistry to geology students. *Journal of Geoscience Education*, 51(5), 481.
- Kockelmans, J. J. (1979). *Interdisciplinarity and higher education*. University Park: Pennsylvania State University Press.
- Krajcik, J., Reiser, B., Fortus, D., & Sutherland, L. (2013). *Investigating and questioning our world through science and technology* (2nd ed.). Greenwich, CT: Sangari Active Science.
- Kuchinke, K. P. (1997). Employee expertises the status of the theory and the literature. *Performance Improvement Quarterly*, 10(4), 72-86.
- Labaree, D. F. (2005). Progressivism, schools and schools of education: An American romance. *Paedagogica Historica*, 41(1-2), 275-288.
doi:10.1080/0030923042000335583
- Lai, E. R., Wei, H., Hall, E. L., & Fulkerson, D. (2012). *Establishing an evidence-based validity argument for performance assessment*. New Jersey: Pearson.
- Lattuca, L. R., Voigt, L. J., & Fath, K. Q. (2004). Does interdisciplinarity promote learning? Theoretical support and researchable questions. *The Review of Higher Education*, 28(1), 23-48.
- Lederman, N. G., & Niess, M. L. (1997). EDITORIAL. *School Science and Mathematics*, 97(2), 57-58. doi:10.1111/j.1949-8594.1997.tb17342.x

- Lee, H. S., & Liu, O. L. (2010). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education*, 94(4), 665-688.
- Lehrer, R., & Schauble, L. (2006). Cultivating model-based reasoning in science education. In R. K. Sawyer (Ed.), *Handbook of the learning sciences* (pp. 371-387). New York, NY: Cambridge University Press.
- Leonard, B. J. B. (2007). *Integrative learning as a developmental process: A grounded theory of college students' experiences in integrative studies*. Unpublished doctoral dissertation. University of Maryland, College Park.
- Linacre, J. M. (2015). *Facets computer program for many-facet Rasch measurement, version 3.71.4*. Beaverton, Oregon: Winsteps.com.
- Linn, M. C. (2006). The knowledge integration perspective on learning and instruction. In R. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences*. Cambridge, MA: Cambridge University Press.
- Linn, M. C., Slotta, J. D., Terashima, H., Stone, E., & Madhok, J. (2010). *Designing science instruction using the web-based inquiry science environment (WISE)*. Paper presented at the Asia-Pacific Forum on Science Learning and Teaching.
- Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-fifth Yearbook of the National Society for the Study of Education* (pp. 84-103). Chicago: University of Chicago.
- Liu, O. L., Lee, H. S., Hofstetter, C., & Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, 13(1), 33-55. doi:10.1080/10627190801968224
- Liu, O. L., Lee, H. S., & Linn, M. C. (2010). An investigation of teacher impact on student inquiry science performance using a hierarchical linear model. *Journal of Research in Science Teaching*, 47(7), 807-819.
- Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 28, 989-1020.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382-386.
- Marzano, R. J. (1991). Fostering Thinking across the Curriculum through Knowledge Restructuring. *Journal of Reading*, 34(7), 518-525.
- McComas, W. F., & Wang, H. A. (1998). Blended science: The rewards and challenges of integrating the science disciplines for instruction. *School Science and Mathematics*, 98(6), 340-348.

- Meeth, L. R. (1978). Interdisciplinary studies: A matter of definition. *Change*, 10(7), 10-10. doi:10.1080/00091383.1978.10569474
- Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.
- Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational measurement: Issues and practice*, 14(4), 5-8.
- Messick, S. (1996). Validity and washback in language testing. *ETS Research Report Series*, 13(3), 241-256.
- Metz, K. E. (1995). Reassessment of Developmental Constraints on Children's Science Instruction. *Review of Educational Research*, 65(2), 93-127.
- Miller, M. D., & Linn, R. L. (2000). Validation of performance based assessments. *Applied Psychological Measurement*, 24, 367-378.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence - centered design. *ETS Research Report Series*, 2003(1), i-29.
- Munier, V., & Merle, H. (2009). Interdisciplinary mathematics–physics approaches to teaching the concept of angle in elementary school. *International Journal of Science Education*, 31(14), 1857-1895.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176.
- Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide*. (7 ed.). Los Angeles, CA: Muthén & Muthén.
- Nagle, B. (2013). Preparing high school students for the interdisciplinary nature of modern biology. *CBE-Life Sciences Education*, 12(2), 144-147.
- National Academy of Sciences. (2004). *Facilitating interdisciplinary research*. Washington, D.C.: National Academies.
- National Assessment Governing Board. (2008). *Science framework for the 2009 national assessment of educational progress*. Washington, DC: NAGB, U.S. Department of Education.
- National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.
- National Research Council. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.
- National Research Council. (2014). *Developing assessments for the next generation science standards* (J. W. Pellegrino, M. R. Wilson, J. A. Koenig, & A. S. Beatty Eds.). Washington, DC: National Academies Press.
- National Research Council (NRC). (2007). Learning progressions. In R. A. Duschl, H. A. Schweingruber, & A. W. Shouse (Eds.), *Taking science to schools. learning and*

- teaching science in grades K-8* (pp. 213-250). Washington, DC: The National Academies Press.
- National Science Teachers Association. (1964). *Theory into action in science curriculum development*. Washington DC: National Science Teachers Association.
- National Science Teachers Association (NSTA). (1998). *Standards for science teacher preparation*. Retrieved from <http://www.nsta.org>.
- Neurath, O. (1996). Unified science as encyclopedic integration. *Logical empiricism at its peak: Schlick, Carnap, and Neurath*, 309-335.
- Newell, W. H. (1994). Designing interdisciplinary courses. *New Directions for Teaching and Learning*, 1994(58), 35-51.
- Newell, W. H., & Green, W. J. (1982). Defining and teaching interdisciplinary studies. *Improving College and University Teaching*, 30(1), 23-30.
- Newell, W. H., & Green, W. J. (1998). Defining and teaching interdisciplinary studies. In W. H. Newell (Ed.), *Interdisciplinarity: Essays from the literature* (pp. 23-34). New York: The College Board.
- NGSS Lead States. (2013). Next generation science standards: For states, by states: National Academies Press Washington, DC.
- Nielsen, M. E. (1989). Integrative learning for young children: A thematic approach. *Educational Horizons*, 18-24.
- Nitko, A. J., & Brookhart, S. M. (2010). *Educational assessment of students* (6 ed.). NJ: Pearson Education.
- Nowacek, R. S. (2005). A discourse-based theory of interdisciplinary connections. *The Journal of General Education*, 54(3), 171-195.
- OECD. (1972). *Interdisciplinarity: Problems of teaching and research in universities*. Paris: OECD.
- OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*: OECD Publishing.
- Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50-64.
- Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27, 289-298.
- Pellegrino, J. W. (2012). Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching*, 49(6), 831-841.
- Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. London: Sage.
- Piaget, J. (1964). Development and Learning. In R. E. Ripple & V. N. Rockcastle (Eds.), *Piaget Rediscovered* (pp. 7-20). New York: Cornell University Press.

- Piaget, J. (1970). Piaget's theory. In P. Mussen (Ed.), *Carmichael's manual of child psychology* (3rd ed., pp. 703-732). New York: Plenum Press.
- Piaget, J. (1978). *The development of thought: equilibration of cognitive structures*. New York: Viking Press.
- Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership*, 56, 8-16.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics*, 4(3), 207-230.
- Reise, S. P., & Waller, N. G. (2002). Item response theory for dichotomous assessment data. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 88-122). San Francisco, CA: Jossey-Bass.
- Reynolds, C. R., & Kaiser, S. M. (1990). Bias in assessment of aptitude. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 611-653). Guilford, New York.
- Rice, J., & Neureither, B. (2006). An integrated physical, Earth, and life science course for pre-service K-8 teachers. *Journal of Geoscience Education*, 54(3), 255.
- Rozin, P. (1976). The evolution of intelligence and access to the cognitive unconscious. In J. M. Sprague & A. A. Epstein (Eds.), *Progress in Psychobiology and Physiological Psychology* (Vol. 6, pp. 245-280). New York: Academic Press.
- Rury, J. L. (1996). Inquiry in the general education curriculum. *The Journal of General Education*, 175-196.
- Sabbag, A. G., & Zieffler, A. (2015). Assessing learning outcomes: An analysis of the goals-2 instrument. *Statistics Education Research Journal*, 14(2).
- Salkind, N. J. (2010). *Encyclopedia of research design*. Thousand Oaks, Calif: Sage.
- Schaal, S., Bogner, F. X., & Girwidz, R. (2010). Concept mapping assessment of media assisted learning in interdisciplinary science education. *Research in Science Education*, 40(3), 339-352.
- Schramm, S. L. (2001). *Transforming the curriculum: Thinking outside the box*. Lanham: R&L Education.
- Shell, D. F., Brooks, D. W., Trainin, G., Wilson, K. M., Kauffman, D. F., & Herr, L. M. (2009). *The unified learning model: How motivational, cognitive, and neurobiological sciences inform best teaching practices*. Berlin, Heidelberg: Springer Science & Business Media.
- Shen, J., Liu, O. L., & Sung, S. (2014). Designing interdisciplinary assessments in sciences for college students: An example on osmosis. *International Journal of Science Education*, 36(11), 1773-1793. doi:10.1080/09500693.2013.879224
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420.

- Simon, H., & Chase, W. (1973). Skill in chess. *American Scientist*, 61(4), 394-403.
- Singh, K., Granville, M., & Dika, S. (2002). Mathematics and science achievement: Effects of motivation, interest, and academic engagement. *The Journal of Educational Research*, 95(6), 323-332.
- Sireci, S. G., Thissen, d., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3), 237-247.
- Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, 2(3), 281-311.
- Smith, J. L., Deemer, E. D., Thoman, D. B., & Zazworsky, L. (2014). Motivation under the microscope: Understanding undergraduate science students' multiple motivations for research. *Motivation and Emotion*, 38(4), 496-512.
- Stark, J. S., & Lattuca, L. R. (1997). *Shaping the college curriculum: Academic plans in action*. San Francisco, CA: Jossey-Bass.
- Stember, M. (1991). Advancing the social sciences through the interdisciplinary enterprise. *The Social Science Journal*, 28(1), 1-14.
- Stichweh, R. (2003). Differentiation of scientific disciplines: causes and consequences. *Unity of Knowledge in Transdisciplinary Research for Sustainability*, 1, 1-8.
- Stock, P., & Burton, R. J. (2011). Defining terms for integrated (multi-inter-trans-disciplinary) sustainability research. *Sustainability*, 3(8), 1090-1113.
- Sweller, J., Van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251-296.
- Taber, K. S. (2005). Conceptual integration and science learners-do we expect too much?
- Taber, K. S. (2008). Exploring conceptual integration in student thinking: Evidence from a case study. *International Journal of Science Education*, 30(14), 1915-1943.
- Tanner, D. (1989). A brief historical perspective of the struggle for an integrative curriculum. *Educational Horizons*, 68(1), 6-11.
- Teresi, J. A., Oceppek-Welikson, K., Ramirez, M., Kleinman, M., Ornstein, K., & Siu, A. (2015). Evaluation of measurement equivalence of the Family Satisfaction with the End-of-Life Care in an ethnically diverse cohort: Tests of differential item functioning. *Palliative Medicine*, 29(1), 83-96. doi:10.1177/0269216314545802
- Tuysuz, M., Bektas, O., & Geban, O. (2014). *Investigating Pre-service Physics and Chemistry Teachers' Conceptual Integration between Physics and Chemistry*. Paper presented at the National Association For Research in Science Teaching, Pittsburgh, PA.
- Vanides, J. T., Yin, Y., Tomita, M., & Ruiz-Primo, M. (2005). Using concept maps in the science classroom. 28(8), 27-31.
- Villaverde, L. E. (2003). *Secondary schools: A reference handbook*. Santa Barbara, CA: ABC-CLIO Inc.

- Weingart, P. (2010). A short history of knowledge formations. In R. Frodemann, K. J. Thomson, & C. Mitcham (Eds.), *The Oxford handbook of interdisciplinarity*. (pp. 3-14). Oxford: Oxford University Press.
- Wiggins, G. P., & McTighe, J. (1998). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum.
- Wilson, M. (2005). *Constructing Measures : An Item Response Modeling Approach*. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Wilson, M. (2008). Cognitive diagnosis using item response models. *Zeitschrift für Psychologie/Journal of Psychology*, 216(2), 74.
- Wolfe, C. R., & Haynes, C. (2003). Assessing interdisciplinary writing. *Peer Review*, 6(1), 126-169.
- Wraga, W. G. (1996). A century of interdisciplinary curricula in American schools. In P. S. Hlebowitsh & W. G. Wraga (Eds.), *Annual Review of Research for School Leaders* (pp. pp. 117-145). New York: Scholastic/National Association of Secondary School Principals.
- Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, 26(3), 171.
- Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Unpublished doctoral dissertation. University of California. Los Angeles.
- Zhang, D., & Crawford, B. (2014). *Learning and teaching crosscutting concepts from cognitive perspectives*. Paper presented at the National Association for Research in Science Teaching (NARST), Pittsburgh, PA.
- Zwickle, A., M. Koontz, T., M. Slagle, K., & T. Bruskotter, J. (2014). Assessing sustainability knowledge of a student population: Developing a tool to measure knowledge in the environmental, economic and social domains. *International Journal of Sustainability in Higher Education*, 15(4), 375-389.